UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

TESIS DOCTORAL

NUEVOS MODELOS MULTIVARIANTES EN LA MEDICIÓN DEL RIESGO CARDIOVASCULAR

AUTOR:
MARTA MARCOS HIDALGO

DIRECTORES:
Mª PURIFICACIÓN GALINDO VILLARDÓN
FERNANDO MALLO FERNÁNDEZ

2015

Nuevos Modelos Multivariantes en la Medición del Riesgo Cardiovascular

---

Memoria que para optar al Grado de Doctor,
por el Departamento de Estadística de la
Universidad de Salamanca, presenta:

Marta Marcos Hidalgo

Salamanca, 2015

Mª PURIFICACIÓN GALINDO VILLARDÓN

Profesora Titular del Departamento de Estadística de la Universidad de Salamanca

FERNANDO MALLO FERNÁNDEZ

Profesor Asociado del Departamento de Dirección y Economía de la Empresa de la Universidad de León

CERTIFICAN:

Que Marta Marcos Hidalgo, Licenciada en Matemáticas, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: **Nuevos Modelos Multivariantes en la Medición del Riesgo Cardiovascular**, y para que conste, firma el presente certificado en Salamanca, a 23 de Noviembre de 2015.

Dra. Mª PURIFICACIÓN GALINDO VILLARDÓN **Dr. FERNANDO MALLO FERNÁNDEZ**

Cuando menos lo esperamos la vida nos coloca delante un desafío que pone a prueba nuestro coraje y nuestra voluntad de cambio

Paulo Coelho

AGRADECIMIENTOS

En primer lugar me gustaría agradecerle a Purificación Galindo su apoyo incondicional, su paciencia por todos esos momentos de dudas por los que he pasado, su manera de entender mi situación personal y saber esperar siempre, por esa motivación que es capaz de transmitirme con sus palabras y por ese ahora o nunca …que tanto tengo y tendré que agradecer. Mil gracias de todo corazón, Puri.

Agradecerle también a Fernando Mallo la dedicación que ha tenido desde un principio con este trabajo, que se inició cuando él concluía su Tesis y que ha sido fuente de inspiración constante en este proyecto. Gracias Fernando por tu tiempo, por transmitirme también esa motivación necesaria para continuar y esa energía para superar las dudas que iban surgiendo por el camino.

A mi familia por pensar que siempre puedo con todo, porque ese pensamiento me ayuda a conseguir lo inimaginable.

A Beatriz por insistirme en cada momento que este esfuerzo merecería la pena, que no podía desistir en el intento porque el futuro es incierto.

A esa ventana al mar que se ha abierto en mi vida y que ha sido fuente de inspiración y motivación constante.

Por último, un agradecimiento general a todas esas personas que han contribuido de alguna manera a que esta Tesis haya podido ser desarrollada.

Contenido

1. INTRODUCCIÓN

1.1. ESCENARIO ACTUAL DE LA ESTIMACIÓN DEL RIESGO CARDIOVASCULAR

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte en todo el mundo según la Organización Mundial de la Salud. Cada año mueren más personas por ECV que por cualquier otra causa. Se calcula que en 2012 murieron por esta causa 17,5 millones de personas, lo cual representa un 31% de todas las muertes registradas en el mundo. De estas muertes, 7,4 millones se debieron a la cardiopatía coronaria, y 6,7 millones, a los accidentes vasculares cerebrales. La mayoría de las ECV pueden prevenirse actuando sobre factores de riesgo, como el consumo de tabaco, las dietas malsanas y la obesidad, la inactividad física o el consumo nocivo de alcohol, utilizando estrategias que abarquen a toda la población.

En los países occidentales las ECV son la principal causa de muerte y una importante fuente de discapacidad, lo que supone a su vez una enorme carga en términos de costes sanitarios. Es por ello que la prevención primaria ha supuesto un objetivo prioritario durante las últimas décadas para gran parte de los países desarrollados, dedicando recursos e implementando estrategias preventivas encaminadas a identificar a aquellos sujetos sanos con mayor riesgo de padecer la enfermedad con el fin de reducir su riesgo.

En nuestro país, según el informe publicado por el Instituto Nacional de Estadística que hace referencia a las causas de muerte del año 2013 y cuyos resultados en enfermedad cardiovascular han sido analizados por la Sociedad Española de Cardiología, uno de cada tres fallecimientos que se produce en nuestro país (el 30,09% respecto al total de defunciones) se debe a las enfermedades del sistema circulatorio, lo que las sitúa como primera causa de muerte por encima del cáncer y de las enfermedades respiratorias.

Por sexos, la mujer española fallece por esta causa casi un 9% más que el hombre, una brecha que ha aumentado en un punto y medio respecto al año anterior

Por comunidades autónomas, Galicia, Andalucía y Asturias son las que cuentan con un porcentaje de mortalidad por causa cardiovascular más elevado; mientras que Canarias, Madrid y País Vasco son las de menor mortalidad cardiovascular.

¿Qué se entiende por enfermedad cardiovascular?

Se entiende por enfermedad cardiovascular(1) cualquier proceso de índole vascular, incluyendo las cardiopatías congénitas, valvulopatías, endocarditis y vasculitis. En sentido estricto deberían incluirse como ECV los procesos de índole aterosclerótico más prevalentes por su interés epidemiológico y preventivo, como son:

- La enfermedad coronaria: afectación de arterias coronarias manifestada por infarto agudo de miocardio, angor pectoris, insuficiencia cardiaca y muerte súbita de origen coronario. En países desarrollados la tercera parte de las muertes y el 50% de las de origen cardiovascular son atribuibles directamente a cardiopatía isquémica.
- La enfermedad cerebrovascular: afectación de arterias carótidas, cerebrales y vertebrales asintomáticas o manifestada por ictus o ataques transitorios.
- La enfermedad vascular periférica: afectación de arterias iliacas y femorales manifestada por clínica de claudicación intermitente o gangrena.
- La aterosclerosis aórtica y los aneurismas aórticos (torácicos y abdominales).

Se define el **riesgo cardiovascular global** como la probabilidad de sufrir un evento cardiovascular en un periodo de tiempo determinado que normalmente es de 5 – 10 años. La estimación del riesgo cardiovascular tiene entre otras la utilidad de identificar a los pacientes de alto riesgo en prevención primaria y ayudar en la toma de decisiones para la intervención con fármacos en la hipertensión arterial y la hipercolesterolemia. Además permite una asignación de los recursos en función de las necesidades, entendiendo como tales el riesgo de sufrir una enfermedad cardiovascular.

Desde la década de los 50 hasta la actualidad se han realizado numerosos estudios encaminados a identificar los **factores de riesgo** que desencadenan la ECV con el fin de poder actuar globalmente sobre ellas de manera más eficaz, permitiéndonos a su vez identificar aquellos sujetos más expuestos a la enfermedad. La expresión factor de riesgo cardiovascular fue acuñada por Jeremiah Stamler y Joseph T. Doyle en 1963 y se define como un rasgo medible o una característica de un individuo que predice la probabilidad de desarrollar una enfermedad manifiesta.

La prevención cardiovascular continúa siendo uno de los grandes retos de nuestra sociedad, ya que este grupo de enfermedades genera una gran morbimortalidad(2). Básicamente, hay dos tipos de estrategias de prevención: la poblacional y la de

individuos de alto riesgo(3). La estrategia poblacional se basa en la implantación de medidas que afectan a toda la población como por ejemplo, la legislación para regular el consumo de tabaco en lugares públicos. La estrategia de alto riesgo se fundamenta en la identificación de aquellos individuos con un riesgo elevado de presentar una enfermedad cardiovascular y la implantación de medidas preventivas individuales según el nivel de riesgo. Para identificar a estos individuos en prevención primaria, se suele utilizar un cribado oportunista y se determinan los factores de riesgo cardiovascular a toda persona que consulte con el sistema sanitario. Para convertir estos factores en estimación del riesgo cardiovascular, hay diferentes funciones o tablas de riesgo.

La necesidad de prever el riesgo de una enfermedad cardiovascular antes de que comience a desarrollarse y ser capaces de conocer y valorar las probabilidades del riesgo ha generado el desarrollo de diversos sistemas de cuantificación o estratificación del riesgo. En líneas generales podemos dividir los sistemas de cuantificación del riesgo en dos grupos:

- Los **sistemas cualitativos** (o semicuantitativos) se determina si el paciente tiene un riesgo bajo, intermedio, alto o muy alto.

- Los **sistemas cuantitativos** facilitan un valor numérico que denominaremos **riesgo cardiovascular**, y que representa la probabilidad de padecer una enfermedad cardiovascular en un determinado período de tiempo, habitualmente 10 años, en una población determinada.

El estudio pionero y referencia mundial que se interesó por el cálculo del riesgo cardiovascular es el *Framingham Heart Study*, iniciado en 1948 por el Servicio de Salud Pública de Estados Unidos. En ese momento poco se conocía sobre las causas de la ECV, pero ya en aquel momento se había constatado la envergadura del problema y la necesidad de tener un conocimiento más profundo para tratar de tener éxito en las estrategias preventivas.

Desde los primeros estudios de Framingham sabemos que la etiología de la ECV es claramente multifactorial, lo que provoca que tanto para su abordaje terapeútico como para la identificación de los sujetos con mayor riesgo, sea necesario tener en cuenta varios factores de riesgo del sujeto que actúan de manera conjunta y cuyo efecto global no puede analizarse de manera aislada sin tener en cuenta el contexto del resto de

factores. Es por ello que en la actualidad las distintas sociedades científicas para el estudio de estas ECV proponen en sus recomendaciones un abordaje integral de todos los factores de riesgo desde una **perspectiva multivariante** y no la valoración aislada de cada uno de ellos de cara a priorizar las actuaciones preventivas.

En Europa y debido a las limitaciones en la extrapolación de los resultados obtenidos del estudio de Framingham para nuestra sociedad, se han desarrollado modelos de estimación de riesgo a partir de datos propios de morbimortalidad. Uno de ellos, es un estudio prospectivo llevado a cabo en Alemania y que se conoce con las siglas de PROCAM(4) (Prospective Cardiovascular Münster) y que fue iniciado en 1978 por el Institute of Arteriosclerosis Research de la Universidad de Münster (Alemania).

A nivel europeo el principal estudio es el que hoy en día se conoce con el nombre de *European Heart Score* y que inicialmente se denominaba SCORE (*Systematic COronary Risk Evaluation*) que publica sus primeros resultados en 2003. En este proyecto se reunieron bases de datos comunes de 12 estudios de cohortes europeos, incluyendo España. Se diseñaron dos modelos diferenciando alto y bajo nivel de riesgo según la incidencia de la ECV en el país de origen.

A partir de las tablas de riesgo actuales es posible llevar a cabo el cálculo de la probabilidad de presentar un evento cardiovascular desde las siguientes perspectivas:

- Estimación del riesgo absoluto.
- Estimación del riesgo relativo de sufrir la enfermedad comparando el riesgo absoluto del individuo con el riesgo que presenta el grupo de la población de su edad y su sexo y con factores de riesgo óptimos. Es una de las interpretaciones recomendadas para personas jóvenes que, aunque expuestas a factores de riesgo cardiovascular, tienen un riesgo bajo por el gran peso de la edad en el cálculo del riesgo.
- La estimación de la edad vascular. La edad vascular es la edad a la que una persona con los factores de riesgo a nivel óptimo alcanzaría el riesgo que actualmente presenta el paciente. También es útil en personas jóvenes expuestas pero con riesgo bajo.

En España se ha publicado recientemente (2015) un modelo de estimación del riesgo cardiovascular obtenido a partir de varias cohortes genuinamente españolas. Este

sistema de puntuación del riesgo se conoce con el nombre de ERICE – Score(5). Dada su reciente publicación no hay estudios comparativos con el resto de sistemas de puntuación del riesgo utilizados en nuestro país.

1.2. OBJETIVOS DE LA TESIS DOCTORAL

En el presente trabajo nos ocuparemos de los sistemas cuantitativos de estimación de riesgo cardiovascular. El objetivo principal de esta Tesis es el de formular la construcción de un sistema de calificación del riesgo cardiovascular que sea capaz de predecir el nivel de riesgo cardiovascular, clasificar a nuevos pacientes susceptibles de recibir una valoración de su riesgo, calificar la salud cardiovascular y que además sea fácilmente interpretable de cara a su uso en prevención primaria. Además el modelo propuesto debe resolver las limitaciones que presentan los sistemas de cuantificación actuales.

La consecución de este objetivo general se plantea a partir de los siguientes objetivos específicos:

1. Analizar los sistemas de predicción del riesgo cardiovascular utilizados actualmente en prevención primaria en España.
2. Estudiar los modelos de estimación del riesgo empleados en la estimación de los sistemas de cuantificación del riesgo detectados a partir del objetivo anterior.
3. Detectar las limitaciones de los sistemas de predicción actuales, tanto las originadas por el modelo de estimación empleado como cualesquiera otras de naturaleza diferente.
4. Plantear y formalizar nuevos modelos de predicción del riesgo cardiovascular que resuelvan las limitaciones de los actuales sistemas. Los nuevos modelos se caracterizarán por ser fácilmente interpretables y capaces de predecir, clasificar y calificar nuevos pacientes.
5. Diseñar un algoritmo general de construcción de un sistema de cuantificación del riesgo cardiovascular y plantear la metodología específica para cada una de sus fases de construcción.

1.3. ESQUEMA DE CONTENIDOS

Esta Tesis está estructurada en tres partes diferenciadas de contenidos. Una primera parte formada por el Capítulo 1 que se plantea como introductoria al escenario actual de los sistemas de cuantificación del riesgo cardiovascular. La segunda parte, que engloba los cuatro capítulos siguientes, pretende estudiar en profundidad los sistemas de medición del riesgo que se utilizan en la actualidad en prevención primaria en España. Igualmente se exponen las limitaciones detectadas de estos sistemas de riesgo cardiovascular. La tercera parte, que se compone de dos capítulos, plantea la nueva propuesta de sistema de predicción del riesgo cardiovascular a partir de un modelo novedoso para la estimación del riesgo en este contexto médico.

A continuación, se especifica cada uno de los capítulos que conforman la estructura del trabajo:

Capítulo 1: Se plantea con el objetivo de introducir el escenario actual de los sistemas de valoración del riesgo cardiovascular. En esta parte además se plantean los objetivos perseguidos en el trabajo.

Capítulo 2: Se estudia el sistema cuantitativo pionero y de referencia mundial que constituye un pilar básico en el estudio de los riesgos de enfermedad cardiovascular: *Framingham Heart Study*. Se expondrán los distintos modelos de estimación del riesgo cardiovascular que se han planteado a lo largo de los años en este estudio, desde el primer modelo publicado en 1991 hasta el más reciente cuya publicación está fechada en 2009. A continuación, se presentan las dos calibraciones que existen de estas tablas de estimación del riesgo para adaptarlas a la realidad epidemiológica de nuestro país. Estas calibraciones se obtienen a partir de dos estudios sobre la población española, *REGICOR* y *DORICA*.

Capítulo 3: Se estudia el proyecto *Systematic COronary Risk Evaluation*, proyecto europeo conocido actualmente con el nombre de *European Heart Score*. Este sistema surgió ante las recomendaciones de las sociedades científicas de realizar las estimaciones del riesgo a partir de cohortes poblacionales propias, ya que se consideraba que en general el modelo de Framingham sobreestimaba el riesgo absoluto de enfermedad cardiovascular cuando se aplicaba en países europeos.

Capítulo 4: Se estudian los modelos estadísticos a través de los que se han elaborado las tablas de predicción del riesgo cardiovascular utilizadas actualmente en nuestro país. Estos modelos de uso extendido en Análisis de Supervivencia, se corresponden con el *Modelo de Riesgos Proporcionales de Cox* y el que manteniendo su terminología inglesa se conoce con el nombre de *Accelerated Failure Time Regression Model*.

Capítulo 5: Se presentan las limitaciones detectadas en los sistemas de riesgo cardiovascular utilizados en España. Se clasifican distinguiendo entre aquellas inherentes a la metodología empleada y otras limitaciones. Estas otras limitaciones no tienen una clasificación tan clara, pero en definitiva son limitaciones encontradas que restan calidad e introducen sesgos en las predicciones realizadas.

Capítulo 6: Se propone el nuevo modelo de estimación del riesgo cardiovascular desde la perspectiva de estimación de una probabilidad asociada a una variable respuesta binaria indicadora de la presencia o no del evento cardiovascular valorado a partir de los factores de riesgo considerados. La metodología propuesta es cuando menos novedosa en los sistemas de cuantificación del riesgo cardiovascular puesto que se propone la estimación a partir de Modelos Logísticos Lineales Híbridos, que constituyen una extensión natural de los modelos logísticos lineales a los que se añade la capacidad para recoger la no linealidad de las variables explicativas.

Capítulo 7: Este capítulo está dedicado a la construcción de un sistema de cuantificación del riesgo cardiovascular a través de los Modelos Logísticos Lineales Híbridos por Expansiones Lineales de Funciones de Base. Este sistema que se diseña como un método teórico general para la estimación del riesgo cardiovascular, se denomina ***Spanish Cardiovascular Risk Scorecard***. Se plantean las distintas fases de construcción del sistema de predicción del riesgo. Y se detalla la metodología a seguir en cada una de ellas con la finalidad de obtener un modelo interpretable, con el triple objetivo de predecir, calificar y clasificar nuevos pacientes susceptibles de recibir una valoración de su riesgo cardiovascular.

2. FRAMINGHAM HEART STUDY

Los sistemas cuantitativos más utilizados en España se basan en dos estudios, el mencionado estudio norteamericano *Framingham Heart Study* que ha propuesto múltiples sistemas de cuantificación del riesgo y el modelo europeo *SCORE Project* (Systematic Coronary Risk Evaluation). En este capítulo analizaremos el modelo norteamericano de Framingham desde su planteamiento inicial hasta los últimos modelos que se obtienen a partir de este estudio.

En 1948 Thomas Royle Dawber diseña el más importante de los estudios epidemiológicos realizados en el análisis de la enfermedad cardiovascular y que denominó *Framingham Heart Disease Epidemiology Study*(6). Se inició por iniciativa del Servicio de Salud Pública de Estados Unidos con la finalidad de estudiar la epidemiología y los factores de riesgo de la enfermedad cardiovascular. Desde su inicio ha sido referencia mundial en el estudio de la enfermedad cardiovascular y sus factores de riesgo, y su producción en el campo de la estimación del riesgo cardiovascular ha sido clave para el desarrollo de esta área del conocimiento.

Se eligió la ciudad de Framingham, situada 32 km al oeste de Boston, Massachusetts, porque en ella se había realizado con éxito un estudio de base poblacional sobre la tuberculosis en 1918, y por su proximidad a los principales centros médicos de Boston; la presencia de grandes empresas y el apoyo prestado por la comunidad médica y la sociedad civil que estaban bien informadas y se mostraban muy colaboradoras.(7)

La primera cohorte la formaron 5.209 habitantes sanos, de entre 30 y 60 años de edad, que se incorporaron al estudio en 1948, para la realización de exámenes bianuales que han continuado desde entonces. Cuatro años después de haberse iniciado *Framingham Heart Study*, con 34 casos de infarto de miocardio en la cohorte, los investigadores identificaron el colesterol elevado y la presión arterial alta como factores importantes en el desarrollo de la enfermedad cardiovascular. En los años siguientes, el estudio de Framingham ha contribuido a identificar otros factores de riesgo de enfermedad cardiovascular que en la actualidad se consideran clásicos(8).

En 1971, se seleccionó a 5.124 hijos e hijas (y sus cónyuges) de la cohorte inicial, para su inclusión en el llamado *Offspring Study*. Finalmente, en 2002, un total de 4.095 participantes se incorporaron a la cohorte de tercera generación del estudio.

A partir de los datos obtenidos en el estudio de Framingham se han desarrollado un amplio abanico de ecuaciones de estimación del riesgo cardiovascular de manera global, o del riesgo coronario y cerebrovascular por separado, sirviendo de base para numerosas guías y recomendaciones de sociedades científicas y organismos oficiales.

El estudio de Framingham constituye un pilar básico en el estudio de los riesgos de enfermedad cardiovascular, y en diferentes formas es ampliamente utilizado para la toma de decisiones terapéuticas en base a la estimación de riesgo proporcionada por el modelo al introducir las características de riesgo del paciente. Desde la publicación en 1991 por parte de Anderson et al., de los modelos para estimar el riesgo cardiovascular y coronario, varios modelos más han sido propuestos por los investigadores del estudio de Framingham, como los de Wilson et al. (1998) y Grundy et al., (1999) para el riesgo coronario. Posteriormente, D'Agostino et al. (2008), propone un modelo de estimación del riesgo cardiovascular global.

2.1. MODELOS DE FRAMINGHAM

2.1.1. Modelo clásico de Framingham (Anderson, 1991)

El estudio de Framingham, iniciado en el año 1948, empezó a dar resultados epidemiológicos en 1960, pero fue la publicación del modelo de estimación del **riesgo coronario** de Anderson en 1991(9) lo que supuso llevar al clínico una herramienta de cálculo para aplicar a los enfermos.

Previamente, multitud de modelos paramétricos habían sido empleados para la predicción del riesgo cardiovascular desde una perspectiva multivariante. Truet, Cornfield y Kannel (1967) (10) emplearon el **análisis discriminante** para este propósito. Esta técnica sólo ajusta un hiperplano separador a la muestra mediante el que es posible discriminar entre aquellos individuos que presentarán el problema y los que no lo presentarán. Ese fue el principal motivo por el que el análisis discriminante fue inmediatamente sustituido por otras técnicas, como el modelo de **regresión logística** o el **modelo de regresión de Weibull**.

Walker y Duncan (1967), Abbot and MacGee (1987) y otros autores emplean para tal fin el modelo de regresión logística(11). El cálculo de la probabilidad de riesgo

cardiovascular a través del modelo de regresión logística se realiza a través de la siguiente expresión:

$$P(x_1, x_2, \dots, x_n) = \frac{1}{1 + \exp\{-(b_0 + b_1 x_1 + \ \dots + b_n x_n)\}}$$

siendo $(x_1, x_2, \dots, x_n)$ las variables de riesgo y $(b_0, b_1, \dots, b_n)$ los coeficientes del modelo que deben ser estimados.

A pesar de la simplicidad del modelo a la hora del cálculo de la probabilidad, éste fue sustituido por otros debido a que en el modelo de regresión logística la estimación sólo es posible realizarla para un intervalo de tiempo prefijado, que normalmente es de 5 ó 10 años.

El modelo de regresión de Weibull presenta la misma problemática en cuanto al horizonte temporal, éste sólo debe ser empleado si la predicción se realiza para un periodo comprendido entre 0 y 4 años.

El modelo de Anderson(9) fue revolucionario en su época por emplear por primera vez un modelo paramétrico que supera en capacidad predictiva a los modelos de regresión empleados hasta ese momento. Está basado en una serie de variables que se consideran de riesgo, entre las que se incluyen: edad, sexo, colesterol HDL, colesterol total, presión arterial sistólica, hábito tabáquico (sí/no), diabetes mellitus (sí/no) e hipertrofia ventricular izquierda (sí/no). La estimación del riesgo se realiza mediante el modelo paramétrico ***Standard Accelerated Failure Time Model***(12), suponiendo que el tiempo de supervivencia se distribuye según una distribución de Weibull. El riesgo estimado está definido como riesgo coronario (fatal o no fatal) a los 10 años e incluye: angina, infarto agudo de miocardio o enfermedad coronaria.

El modelo de Anderson de 1991, sólo puede emplearse en prevención primaria, ya que en el caso de prevención secundaria los riesgos serían más elevados y las estimaciones ya no tendrían validez. Además, y dado que la ecuación de Framingham está basada en una población de estudio norteamericana de alto riesgo cardiovascular, tiene únicamente validez para aquellas poblaciones de riesgo elevado similar a la población original del estudio.

2.1.2. Modelo de Framingham por categorías de Wilson (Wilson, 1998)

El modelo de Framingham por categorías de Wilson(13) trata de ver la relación entre las categorías de hipertensión arterial, colesterol HDL y LDL del sexto Informe del *Joint National Commitee* de 1997 y del *National Cholesterol Education Program* de 1994, con el riesgo de presentar una enfermedad coronaria (angina estable, inestable, infarto agudo de miocardio (IAM) y muerte coronaria), en un periodo de 10 años.

El modelo de Framingham por categorías de Wilson, analiza edad (30 – 74 años), sexo, presión arterial sistólica y diastólica, colesterol HDL, colesterol LDL y las variables diabetes mellitus y tabaquismo a un total de 2489 hombres y 2856 mujeres a los que se les hizo el seguimiento durante doce años.

La estimación del riesgo se realiza mediante el M**odelo de Riesgos Proporcionales de Cox**(14). Se elaboran tablas para la predicción del riesgo coronario a diez años distinguiendo por sexo y considerando las variables colesterol total (TC) y colesterol LDL como variables categóricas. Se detecta como significativa la interacción entre TC y edad aunque no se incluye en el modelo. El término que sí se incluye en el caso de las mujeres es un término cuadrático para la variable edad por ser éste significativo.

2.1.3. Modelo de Framingham de Grundy (1999)

El modelo de Framingham de Grundy (15) adapta el modelo anterior para estimar el riesgo de enfermedad isquémica grave, permitiendo el cálculo del riesgo de presentar *"hard CHD"*, que incluye sólo la angina inestable, IAM y muerte coronaria. Además, este modelo facilita tanto el *riesgo absoluto* como el *riesgo relativo*.

2.1.4. Modelo de Framingham de D'Agostino (2000)

El Modelo de Framingham D'Agostino(16) fue novedoso en su época por calcular el riesgo coronario no sólo en prevención primaria sino que también en prevención secundaria. El modelo empleado para el cálculo del riesgo es el ***Standard Accelerated Failure Time Model***(12), suponiendo que el tiempo de supervivencia se distribuye según una distribución de Weibull, con un horizonte temporal de cuatro años.

Este modelo de D'Agostino utiliza un mayor número de factores de riesgo, lo que complica su aplicación en la práctica clínica. Las variables utilizadas en prevención primaria son: edad, colesterol total y colesterol HDL, diabetes (sí/no), hábito tabáquico (sí/no), presión arterial sistólica (tratada/no tratada), ingesta de alcohol y en el caso de las mujeres el estado post – menopaúsico (sí/no).

El tratamiento de los factores de riesgo como variables continuas, al contrario que otros modelos como el de categorías de Wilson, mejora las predicciones del riesgo cardiovascular. Incluye en ambos sexos la ratio colesterol total/colesterol HDL con el fin de reducir los efectos de la asimetría de la distribución de la variable y para evitar el incluir los términos de interacción con la variable edad.(16)

2.1.5. Modelo de riesgo cardiovascular general de D'Agostino (2008)

El modelo de riesgo cardiovascular general de D'Agostino(17) está enfocado para uso en atención primaria, primando la simplicidad e incluyendo factores de riesgo clásicos. Este modelo evalúa el riesgo cardiovascular global a diez años mediante el **Modelo de Riesgos Proporcionales de Cox**(14) diferenciado ambos sexos. Las variables que emplea son: edad, colesterol HDL, colesterol total, tensión arterial sistólica (tratada o no), fumador (sí/no) y diabetes (sí/no).

2.1.6. Modelo de riesgo cardiovascular de Framingham a 30 años (2009)

El modelo estimación del riesgo de Framingham a 30 años(18) surge ante el interés manifestado por muchos expertos de conocer el riesgo cardiovascular para un horizonte temporal superior a 10 años, margen habitual hasta esa fecha.

En un primer modelo, se evaluó el efecto de los factores de riesgo en la estimación del riesgo cardiovascular denominado *"hard"* CVD (*Cardiovascular Disease*), incluyendo esta terminología *"hard"* CHD (muerte coronaria, infarto de miocardio) e infarto cerebral (fatal y no fatal). El modelo empleado para tal estimación fue el **Modelo de Regresión de Cox**(14). En un segundo modelo, se evaluó el riesgo de presentar un evento cardiovascular global siguiendo la terminología adoptada por D'Agostino(17).

Según la descripción de la evolución de los modelos de estimación del riesgo cardiovascular derivados del estudio de Framingham, desde sus primeros modelos en 1991 hasta el último publicado en 2009, se han modificando cuestiones relativas a:

- Las variables empleadas en el modelo (edad, sexo, colesterol total, colesterol CHL, diabetes mellitus, presión arterial sistólica…).
- El riesgo que se evalúa, siendo éste en unos casos riesgo exclusivamente de presentar enfermedad coronaria y en otros casos riesgo cardiovascular global.

En cuanto al modelo estadístico empleado para realizar la estimación del riesgo cardiovascular, el modelo de Wilson (1998), el de Framingham D'Agostino (2008) y Framingham a 30 años (2009) implementan el Modelo de R*iesgos Proporcionales de Cox*(14); mientras que el modelo clásico de Anderson (1991) y el de Framingham D'Agostino (2000) emplean el modelo *Accelerated Failure Time*(12) suponiendo que la variable aleatoria tiempo de supervivencia se distribuye según una distribución Weibull.

2.2. CALIBRACIÓN PARA LA POBLACIÓN ESPAÑOLA DE LOS MODELOS DE FRAMINGHAM

Las tablas de riesgo están diseñadas para una población con características específicas por lo que su uso fuera de estas poblaciones requiere previamente de una adaptación a la realidad epidemiológica de la población en la que se espera introducir, es decir, requieren de una calibración previa ya que de no hacerlo pueden sobreestimar o subestimar el riesgo cardiovascular de los individuos a los que se les aplica.

En el caso de las tablas de riesgo de Framingham, diversos estudios han puesto de manifiesto que la diferencia en la incidencia de acontecimientos coronarios y en la prevalencia de factores de riesgo hacen que su aplicación en poblaciones con baja incidencia como la mediterránea sobreestime el riesgo(19). Este hecho es debido a que las predicciones de los modelos de riesgo son fuertemente dependientes de la población de origen de la que se han extraído los datos, y por tanto no son extrapolables directamente los resultados de un modelo de riesgo como es el estudio de Framingham basado en una cohorte de una población norteamericana a la población española, ya que

la incidencia de enfermedad coronaria española es inferior a la de la población de origen.

Por ello y tratando de dar respuesta a la Tercera Recomendación Europea conjunta de prevención cardiovascular(20) en la que se indica que las tablas deben ser adaptadas recogiendo el nivel de riesgo y las tasas de mortalidad de cada país, el trabajo publicado en 2003 bajo el título *"Estimación del riesgo coronario en España mediante la ecuación de Framingham calibrada"*(21) presentan las tablas de riesgo calibradas para nuestro país a partir del registro poblacional de infarto de miocardio de Girona REGICOR (Registre Gironí del Cor)(22). Posteriormente otros estudios como el publicado en 2004 bajo el título *"Tablas de evaluación del riesgo coronario adaptadas a la población española. Estudio DORICA"*(23) calibran las tablas originales mediante los datos obtenidos a partir de un estudio epidemiológico nutricional y de factores de riesgo cardiovascular de carácter transversal, realizados entre 1990 y 2000.

La calibración de la ecuación consiste en la sustitución del elemento de comparación promedio de Framingham por uno local(24). Para ello, es necesario disponer de una estimación fiable de la prevalencia local de los factores de riesgo involucrados, de la tasa local de incidencia de los acontecimientos coronarios considerados, y de los coeficientes originales de la ecuación.

La probabilidad de un acontecimiento coronario en un tiempo t, en un paciente con un conjunto de factores de riesgo x_i , $1 \leq i \leq p$, se expresa mediante:

$$P(t) = 1 - S_0(t)^{exp(\sum_{i=1}^{p} \beta_i x_i - \sum_{i=1}^{p} \beta_i \bar{x}_i)}$$

siendo:

$S_0(t)$ la probabilidad basal de estar libre de acontecimientos coronarios en el tiempo t.

$\sum_{i=1}^{p} \beta_i \bar{x}_i$, función lineal de riesgo promedio en el conjunto de valores $\bar{x}_i$, $1 \leq i \leq p$, de cada grado de cada factor en la población de referencia.

$\sum_{i=1}^{p} \beta_i x_i$, la función lineal calculada en el conjunto de valores x_i , $1 \leq i \leq p$ que presenta el valor de cada factor en el individuo objeto de estudio.

β_i , $1 \leq i \leq p$, coeficientes de la función de riesgos proporcionales de Cox para cada grado de cada factor considerado.

De las calibraciones que se han llevado a cabo de las tablas de Framingham cabe destacar fundamentalmente la realizada mediante el estudio de *REGICOR* y la basada en el estudio *DORICA* que describimos a continuación.

2.2.1. Calibración basada en el estudio REGICOR

En 2003 y en ausencia de estudios poblacionales de cohorte, Marrugat *et al*.(21), publican la adaptación de la ecuación de riesgo coronario de Framingham calibrada para la población española, a partir de los datos del registro poblacional de infartos de miocardio de Gerona REGICOR (Registre Gironí del Cor)(22).

Para ello sustituyeron en las ecuaciones de Framingham Wilson (1998) la prevalencia de los factores de riesgo cardiovascular y la tasa de incidencia de eventos coronarios de Framingham por los de nuestro medio, que en este estudio se consideraron los obtenidos a partir del registro REGICOR. La tasa de acontecimientos mayores se obtiene a partir de los datos de REGICOR, que investiga todos los casos sospechosos de IAM en seis comarcas de Girona. La tasa de incidencia de IAM silente y de angina eran desconocidas en Gerona, por ello se asumió que la proporción era similar a la de Framingham(21). Este aspecto confiere a la nueva tabla calibrada un carácter conservador, ya que es poco probable que los valores reales de Girona sean superiores a los de la ciudad americana.

Una de las principales limitaciones de esta calibración se fundamenta en el hecho de que la incidencia de IAM en Girona se encuentra aproximadamente un 15% por debajo del promedio de España según el estudio IBERICA (Investigación y Búsqueda Específica y Registro de Isquemia Coronaria Aguda)(25). Por tanto, la validez externa de la ecuación y las tablas que de ella se derivan a otras zonas de España debe ser aceptada con las debidas precauciones.

Un estudio comparativo entre la ecuación de Framingham Wilson(13) y la ecuación calibrada REGICOR(21) demuestra la sobrevaloración que se obtiene al calcular el riesgo mediante la función de Framingham. Según los resultados del estudio implicaría

un mayor porcentaje de pacientes potencialmente tratables con fármacos hipolipemiantes(26). Este hecho apoya la necesidad de disponer de tablas de riesgo cardiovascular ajustadas para nuestra población.

2.2.2. Calibración basada en el estudio DORICA

El estudio DORICA se llevó a cabo a partir de un conjunto de datos configurado por estudios epidemiológicos nutricionales y de factores de riesgo cardiovascular de carácter transversal, realizados entre 1990 y 2000 sobre muestras aleatorias representativas de la población de Andalucía, Baleares, Canarias, Cataluña, Galicia, Madrid, Región de Murcia, País Vasco y Comunidad Valenciana. Para la obtención de la muestra se siguió en todos los casos un procedimiento de muestreo aleatorio polietápico estratificado según la edad, el sexo y el hábitat, por asignación proporcional a la densidad de población. En este estudio se incluyó población adulta con edades comprendidas entre 25 y 64 años ($n = 14.616$, 6.796 varones y 7.820 mujeres).

En 2004, Javier Aranceta et al.(23) publican la adaptación de la ecuación de Framingham calibrada a través del estudio DORICA, y nace como alternativa a la ecuación basada en el estudio REGICOR. Se confeccionan tablas de riesgo coronario partiendo también de la ecuación de Framingham propuesta por Wilson, pero en este caso adaptando la prevalencia de factores de riesgo en la población española a partir de las estimaciones realizadas en el estudio DORICA(23).

3. EUROPEAN HEART SCORE

Las sociedades europeas teniendo en cuenta las limitaciones en la extrapolación de los resultados obtenidos del estudio de Framingham, han desarrollado modelos de estimación de riesgo a partir de datos propios de morbimortalidad.

Uno de ellos, es un estudio prospectivo llevado a cabo en Alemania y que se conoce con las siglas de PROCAM(4) (*Prospective Cardiovascular Münster*), fue iniciado en 1978 por el *Institute of Arteriosclerosis Research* de la Universidad de Münster (Alemania). En este estudio se evaluaron más de 20.060 varones trabajadores de 52 compañías, incluyendo autoridades gubernamentales. Del total de trabajadores evaluados fueron incluidos en la cohorte 5.389 varones de 35 a 65 años. Con la información obtenida de la cohorte se construyó un algoritmo de predicción del "riesgo coronario restringido" ya que al igual que otros modelos como el de *Framingham Grundy* excluye el ángor estable e inestable y la isquemia silente. La estimación se realizó mediante el **Modelo de Riesgos Proporcionales de Cox**(14) considerando los factores de riesgo edad, colesterol LDL, colesterol HDL, niveles de triglicéridos, antecedentes familiares de infarto agudo de miocardio, tensión arterial sistólica, diabetes (sí/no) y fumador (sí/no) que se detectaron entre las 57 variables que midió el estudio, como variables independientes para la medición del riesgo(4).

Uno de los principales inconvenientes de este estudio es el de la elección de la población de partida, exclusivamente varones trabajadores, por lo que la representatividad de la misma en las futuras extrapolaciones no estaría asegurada. En 2002 se ofrecían resultados preliminares para un subgrupo de 2.810 mujeres de 45 a 65 años. Estos resultados posteriormente han sido ampliados proponiendo nuevas ecuaciones para aproximar el riesgo de una manera más exacta a esta población.

Para aplicar este modelo a otros países, habría que aplicar un factor de corrección que tiene en cuenta las características y estadísticas por países, permitiendo así una estimación más fiable del riesgo según las características de la población local. A pesar de ello, la función PROCAM se comporta de manera similar a la de Framingham y sobrestima el riesgo en países mediterráneos, pero también en países de riesgo alto o intermedio(27).

A finales de 2007 se publican nuevos datos del estudio PROCAM con importantes novedades. Amplían la población en estudio a 18.460 hombres y 8.515 mujeres, e introducen en la evaluación del riesgo coronario una nueva ecuación, derivada del

Modelo de Riesgos Proporcionales de Cox suponiendo que el tiempo de supervivencia sigue una distribución de Weibull. Este nuevo modelo permite medir de una forma más exacta el riesgo en mujeres y por vez primera se puede evaluar el riesgo en sujetos mayores de 65 años(28). En este mismo estudio se introduce la evaluación independiente del riesgo de ictus mediante el modelo de Cox.

Las sociedades científicas relacionadas con el riesgo cardiovascular global o coronario en Europa, consideraron desde un principio que en general el modelo de Framingham sobreestimaba el riesgo absoluto de enfermedad cardiovascular cuando se implementaba en países europeos(29).

De esta forma se creó la necesidad de desarrollar un modelo de acuerdo con el verdadero riesgo cardiovascular de estos países. El proyecto, se denominó Proyecto SCORE (*Systematic COronary Risk Evaluation*)(30), y sus resultados son publicados en 2003. Este modelo de estimación hoy en día se denomina *European Heart SCORE* o *EuroSCORE* y es el recomendado por las Sociedades Europeas(31).

Se incluyeron estudios de cohortes europeas con 205.178 individuos (117.098 hombres y 88.080 mujeres) de 24 a 75 años. A pesar de ello, el ajuste del modelo se realiza para el grupo de edad de 45 a 64 años debido a que la edad es el principal determinante de riesgo coronario y los rangos de las cohortes son muy heterogéneos.

Se reunieron bases de datos comunes de 12 estudios de cohortes europeos, incluyendo España que estuvo representada nuevamente por Cataluña. Junto a España se incluyeron Finlandia, Rusia, Noruega, Reino Unido, Escocia, Dinamarca, Suecia, Bélgica, Alemania, Italia y Francia. Dada la variabilidad geográfica del riesgo cardiovascular en Europa(32), se desarrollaron dos modelos SCORE, distinguiendo países de alto y bajo riesgo. Para tipificar las cohortes como población de alto y bajo riesgo se tuvo en cuenta tanto la tasa de muerte por enfermedad cardiovascular como las estadísticas nacionales de mortalidad. Las cohortes de Dinamarca, Finlandia y Noruega, junto con los coeficientes de los factores de riesgo derivados del conjunto de datos, se utilizaron para desarrollar el alto riesgo del modelo, mientras que las cohortes de Bélgica, Italia y España fueron las líneas de base para la función de supervivencia para desarrollar el modelo de la región de bajo riesgo(30).

La novedad más importante de la función de riego SCORE comparada con la de Framingham es que estima el riesgo mortal de todas las manifestaciones aterotrombóticas cardiovasculares, incluidos el ictus, la insuficiencia cardiaca, la insuficiencia arterial periférica o ciertos aneurismas y no sólo la enfermedad coronaria. Por tanto, la aplicación de las tablas SCORE permite la estimación del riesgo de desarrollar en 10 años una **enfermedad cardiovascular fatal** en países de Europa, a partir de una población autóctona. Aunque en un principio se consideró el tratar tanto los eventos cardiovasculares fatales como los no fatales y ECV no coronarios, finalmente se optó por incluir sólo los eventos fatales, con resultado de muerte. No cabe duda que tanto los pacientes como los médicos tienen también interés en los eventos no fatales, y de hecho son la mayor carga económica para el sistema sanitario. El inconveniente está en que los eventos no fatales pueden plantear problemas en el desarrollo del sistema de predicción del riesgo, ya que son criticables dependiendo de las definiciones y métodos usados en su comprobación. Además, no todos los países disponen de estudios de cohortes de enfermedad cardiovascular no fatal, sin embargo todos disponen de un registro de datos referidos a la mortalidad específica por causas. Esta elección de los eventos fatales en el proyecto SCORE, se diferencia de forma clara del estudio de Framingham que incluían en sus objetivos la enfermedad coronaria no fatal, blanda (angina) y dura (SCA, IAM no fatal).

El cálculo del riesgo con un horizonte temporal de 10 años, se realiza mediante el **modelo de riesgos proporcionales de Weibull**, es decir el Modelo de Riesgos Proporcionales de Cox(14) en el que se supone que la distribución de la variable aleatoria tiempo de supervivencia sigue una distribución Weibul.

El tratamiento de la variable edad es uno de los aspectos más novedosos del proyecto SCORE, ya que se considera como una medida del tiempo de exposición al riesgo en lugar de como un factor de riesgo. El enfoque tradicional propuesto en la mayoría de los estudios es aquel en el que la edad se considera como un factor de riesgo, teniendo la desventaja de que la supervivencia no puede ser estimada para tiempos de seguimiento superiores a la duración del periodo de seguimiento del estudio. Sin embargo, en el proyecto SCORE se usa la edad como variable tiempo, lo que permite hacer estimaciones para todo el rango de edad observado en el estudio.

El modelo fue estratificado en cohortes por sexo, obteniéndose modelos de estimación diferentes para hombres y mujeres. El riesgo de muerte cardiovascular se realiza mediante dos estimaciones independientes de riesgo: un modelo para el riesgo de enfermedad coronaria y otro para enfermedad aterosclerótica no coronaria, permitiendo así la estimación del riesgo cardiovascular global.

Aunque la metodología de este modelo la estudiaremos en detalle en el capítulo 4, veamos los cálculos necesarios para la obtención de los valores de riesgo de muerte cardiovascular(30).

1º Probabilidad de supervivencia en el momento actual y dentro de diez años.

$$S_0(edad) = \exp\{-\exp(\alpha)\,(edad - 20)^p\}$$

$$S_0(edad + 10) = \exp\{-\exp(\alpha)\,(edad - 10)^p\}$$

donde los valores de $\alpha\ y\ p$, están tabulados en la siguiente tabla:

		CHD		*Non –CHD CVD*	
		α	p	α	p
Bajo riesgo	Hombres	$-22{,}1$	4,71	$-26{,}7$	5,64
	Mujeres	$-29{,}8$	6,36	$-31{,}0$	6,62
Alto riesgo	Hombres	$-21{,}0$	4,62	$-25{,}7$	5,47
	Mujeres	$-28{,}7$	6,23	-30	6,42

2º Probabilidad de supervivencia a una edad prefijada:

$$S(edad) = \{S_0\,(edad)\}^{exp\{\sum_{i=1}^{2}\beta_i(x_i-\bar{x}_i)+\beta_3 x_3\}}$$

$$S(edad + 10) = \{S_0\,(edad + 10)\}^{exp\{\sum_{i=1}^{2}\beta_i(x_i-\bar{x}_i)+\beta_3 x_3\}}$$

$siendo$:

$x_{i,}\ i = 1, 2$, los valores de los factores de riesgo colesterol y presión arterial.

x_3, variable cualitativa dicotómica referente al hábito de fumar.

$\beta_{i,}\ i = 1, 2{,}3$, los coeficientes del modelo.

3º La probabilidad de supervivencia a diez años a partir de las anteriores y el riesgo a diez años:

$$S_{10}(edad) = \frac{S(edad + 10)}{S(edad)}$$

$$Risk_{10} = 1 - S_{10}(edad)$$

4º Combinamos los riesgos coronarios y no coronarios calculados y obtenemos el riesgo cardiovascular total mediante la suma de ambos:

$$CVDRisk_{10}(edad) = \boldsymbol{CHD}Risk(edad) + \boldsymbol{Non} - \boldsymbol{CHD}Risk(age)$$

El cálculo de la estimación del riesgo puede hacerse de dos formas diferentes según el tratamiento que reciba la variable colesterol, una estimación está basada en el colesterol total y otra hace uso de la razón entre el colesterol total y el colesterol – HDL.

En el proceso de adaptación de las guías de prevención se aconsejó calibrar los modelos *EuroSCORE* al nivel del riesgo de cada país. Las tablas calibradas se calcularon mediante la utilización de las tasas de mortalidad española y los factores de riesgo del estudio MONICA – Catalunya (MONItoring CArdiovascuar Diseases)(33). El modelo *EuroSCORE* calibrado para España produce riesgos superiores en un 13% al de la función SCORE de bajo riesgo, aunque las diferencias entre ambas oscilaron según la edad, el sexo y especialmente el tabaco. Esta calibración reconoce algunas situaciones de riesgo no identificadas por la tabla *EuroSCORE* original, que son clínicamente más obvias; por ejemplo, mujeres no fumadoras de 60 años con una presión arterial de 180 mmHg y colesterol total de 8 mmol/l. En ambas tablas hay muy pocas mujeres de alto riesgo antes de los 60 años. El porcentaje máximo de sujetos nuevamente identificados de alto riesgo con la tabla calibrada fue del 22%, observándose más diferencias en los varones mayores de 55 años.

4. MODELOS MULTIVARIANES ACTUALES EN LA MEDICIÓN DEL RIESGO CARDIOVASCULAR

4.1. INTRODUCCIÓN

En la actualidad, la probabilidad de padecer una enfermedad cardiovascular en un determinado periodo de tiempo se puede estimar a través de las tablas de riesgo o de calculadoras de riesgo cardiovascular. En ambos casos, la estimación se realiza a partir de un determinado modelo multivariante, empleando para la estimación una serie limitada de variables explicativas o factores de riesgo.

En España las funciones de riesgo utilizadas parten de los dos estudios descritos en el Capítulo 2 FRAMINGHAM HEART STUDY y en el Capítulo 3 EUROPEAN HEART SCORE. Las distintas ecuaciones que se plantean para estimar el riesgo cardiovascular en estos estudios se basan en dos de las técnicas comúnmente utilizadas en el análisis de la supervivencia: Modelo de riesgos proporcionales de Cox(14) y *Accelerated Failure Time Regression Model*(12).

En este capítulo se introducirán primeramente una serie de conceptos generales y terminología del Análisis de Supervivencia, que son necesarios para el posterior estudio teórico de los modelos de estimación del riesgo que se detallan. La razón fundamental por la que se lleva a cabo este estudio exhaustivo de los modelos de estimación a partir de los que se elaboran las tablas actuales de estimación del riesgo cardiovascular en España, es la de valorar la calidad de las predicciones que de ellos puedan extraerse y detectar las posibles limitaciones metodológicas de los sistemas de cuantificación que emplean esta metodología estadística.

4.2. CONCEPTOS Y TERMINOLOGÍA

Sea T variable aleatoria absolutamente continua que representa tiempo hasta que ocurre un evento, en nuestro caso particular representará tiempo hasta la primera aparición de un suceso cardiovascular.

Consideremos las siguientes funciones:

Función de distribución acumulada, también denominada distribución de tiempo de vida o de fallo.

$$F_T(t) = P(T \leq t) \qquad (4.1)$$

Función de distribución de supervivencia, tal que $S_T(0) = 1$, puesto que la variable aleatoria T es no negativa.

$$S_T(t) = 1 - F_T(t) = P(T > t) \quad (4.2)$$

Función de densidad de probabilidad asociada, que representa la tasa instantánea absoluta de muerte.

$$f_T(t) = \frac{dF_T(t)}{dt} = -\frac{dS_T(t)}{dt} \quad (4.3)$$

La función de riesgo (*hazard rate function*), que representará la tasa instantánea de fallo en el instante de tiempo t, que la denotamos por $h_T(t)$, se define como:

$$h_T(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \quad , \quad t \geq 0 \quad (4.4)$$

La función de riesgo nos da una idea de la tasa instantánea de riesgo, y también se denomina función de peligro o función de impacto.

A partir de las definiciones anteriores, se tiene que:

$$P(t \leq T < t + \Delta t \mid T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{F_T(t + \Delta t) - F_T(t)}{S_T(t)}$$

$$\Rightarrow h_T(t) = \lim_{\Delta t \to 0} \left(\frac{F_T(t + \Delta t) - F_T(t)}{\Delta t} \right) \cdot \frac{1}{S_T(t)} = \frac{f_T(t)}{S_T(t)}$$

Por tanto, se tiene la siguiente relación entre las funciones definidas:

$$h_T(t) = -\frac{d log S_T(t)}{dt} \quad (4.5)$$

de donde se tiene que la función de riesgo acumulada $H_T(t)$ y la función de supervivencia $S_T(t)$ se relacionan como sigue:

$$H_T(t) = \int_0^t h_T(u)du = -log\ S_T(t)$$

$$S_T(t) = exp\{-\mathrm{H}_T(t)\}$$

La función cuantil, para $0 < p < 1$

$$t_p = inf\{\, t : \ F_T(t) \geq p\}$$

Destacar que si $F_T(t)$ es continua y estrictamente creciente entonces:

$$t_p = F_T^{-1}(p)\,, \qquad 0 < p < 1$$

Para un valor prefijado p, el cuantil t_p expresa el tiempo para el cual una proporción determinada p de la población falla, lo que en nuestro contexto significa que se presenta el evento cardiovascular valorado.

La media y la varianza de la variable aleatoria T son dos características importantes al caracterizar la supervivencia:

$$E(T) = \int_0^\infty S(t)dt \qquad\qquad Var(T) = 2\int_0^\infty tS(t)dt - \{E(T)\}^2$$

4.3. MODELO DE REGRESIÓN DE COX

En 1972 Cox publicó un artículo "*Regression models and life tables"*(14), que es uno de los artículos más citados en la bibliografía científica.

La regresión de Cox se utiliza cuando la variable dependiente está relacionada con la supervivencia de los individuos y se desea averiguar simultáneamente el efecto independiente una serie de factores sobre esta supervivencia.

Sea $X_{n\times p}$ la matriz de datos, en la que las filas representan los n individuos y las columnas las p variables consideradas.

Sean,

$X_{ij}(t)$ el valor de la variable $j-$ésima del individuo $i-$ésimo, siendo $i = 1, \dots, N$ y $j = 1, \dots, p$.

$X_i(t)$ el vector asociado al individuo $i-$ésimo, siendo $i = 1, \dots, N$; también denotado como X_i, puesto que las distintas mediciones de las variables se realizarán en el mismo instante de tiempo, t.

El **Modelo de Cox**(14) define la función de riesgo para el individuo $i-$ésimo como:

$$h_i(t) = h_0(t)\, exp\{\beta_1 X_{i1}(t) + \dots + \beta_p X_{ip}(t)\} = h_0(t) exp\{\beta' X_i(t)\} \qquad (4.6)$$

donde $h_0(t)$ es no negativa y se conoce como "función de riesgo basal", y es aquella función de referencia en la que el valor de todas las variables incluidas en el modelo toman el valor cero, y $\beta' = (\beta_1, \dots, \beta_p)$ es el vector de coeficientes de regresión de Cox.

Por tanto, según (4.6) el Modelo de regresión de Cox es un modelo lineal que está formado por el producto de dos términos. El primero depende exclusivamente del tiempo, mientras que el segundo depende sólo de las variables. El Modelo de Cox se caracteriza por no especificar la forma de $h_0(t)$**.** Se denomina un modelo semiparamétrico porque se estiman los parámetros $\beta_j,\ 1 \leq j \leq p$, mientras que el valor de la función de riesgo de referencia $h_0(t)$ se obtiene a partir de los datos.

La correspondiente función de supervivencia tendrá la siguiente expresión:

$$S_i(t) = S_0(t)^{exp\left\{\sum_{j=1}^{p} \beta_j X_{ij}(t)\right\}} \quad (4.7)$$

El modelo también es conocido como **modelo de riesgos proporcionales**, ya que fijados dos individuos cualesquiera, el cociente de sus funciones de riesgo es independiente del tiempo:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\, e^{X_i\beta}}{\lambda_0(t)\, e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}} \quad (4.8)$$

La estimación de β se realiza a través de la función parcial de verosimilitud introducida por Cox(14):

$$PL(\beta) = \prod_{i=1}^{N} \prod_{t \geq 0} \left\{ \frac{Y_i(t) r_i(\beta, t)}{\sum_j Y_j(t) r_j(\beta, t)} \right\}^{dN_i(t)} \quad (4.9)$$

donde, $r_i(\beta, t)$ es el riesgo asociado al individuo $i - ésimo$, $i = 1, \dots, n$:

$$r_i(\beta, t) = \exp\left[X_i(t)\beta\right] \equiv r_i(t)$$

Considerando logaritmos en la función de verosimilitud de Cox (4.9), se obtiene:

$$l(\beta) = \sum_{i=1}^{N} \int_0^{\infty} \left[Y_i(t) X_i(t) \beta - log\left(\sum_j Y_j(t) r_j(t) \right) \right] dN_i(t)$$

de donde, derivando la función anterior con respecto a β, que denotamos por $U(\beta)$:

$$U(\beta) = \sum_{i=1}^{N} \int_0^{\infty} [X_i(s) - \bar{x}(\beta, s)]\, dN_i(s)$$

siendo $\bar{x}(\beta, s)$ la media ponderada de X :

$$\bar{x}(\beta, s) = \frac{\sum_{i=1}^{N} Y_i(s) r_i(s) X_i(s)}{\sum_{i=1}^{N} Y_i(s) r_i(s)}$$

El estimador de máxima verosimilitud parcial se obtiene resolviendo la ecuación:

$$U(\hat{\beta}) = 0 \qquad (4.10)$$

La solución, $\hat{\beta}$ es consistente y asintóticamente normal con media β, el verdadero valor del parámetro, y varianza $\{E[I(\beta)]\}^{-1}$, es decir la inversa del valor esperado de la matriz de información $I(\beta)$, con:

$$I(\beta) = \sum_{i=1}^{N} \int_0^{\infty} V(\beta, s) dN_i(s)$$

siendo $V(\beta, s)$ la varianza ponderada, con:

$$V(\beta, s) = \frac{\sum_{i=1}^{N} Y_i(s) r_i(s) [X_i(s) - \bar{x}(\beta, s)]' [X_i(s) - \bar{x}(\beta, s)]}{\sum_{i=1}^{N} Y_i(s) r_i(s)}$$

Para resolver la ecuación (4.10) se emplea el algoritmo iterativo de Newton – Raphson. Se comienza con un valor inicial $\hat{\beta}^{(0)}$, y de manera iterativa hasta la convergencia se calcula:

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + I^{-1}\left(\hat{\beta}^{(n)}\right) U(\hat{\beta}^{(n)})$$

Actualmente se disponen de paquetes estadísticos como *SAS* y *S – Plus* para la resolución de dicha ecuación.

A partir del planteamiento del Modelo de Regresión de Cox, se tiene que:

$$h_i(t) = h_0(t) exp\{\beta' X_i(t)\} \Rightarrow ln\left(\frac{h_i(t)}{h_0(t)}\right) = \beta' X_i(t) \qquad (4.11)$$

Por tanto, el modelo de riesgos proporcionales puede considerarse como un modelo lineal para el logaritmo de la ratio de las funciones de riesgo, es decir, plantea el logaritmo del riesgo relativo como un modelo lineal de las variables independientes. Se supone por tanto, que el riesgo relativo a diferencia del riesgo propiamente dicho, no depende del tiempo o, dicho de otra manera, que es constante a lo largo del tiempo.

El Modelo de Regresión de Cox, que supone que para cada variable el riesgo relativo (*hazard ratio*) es constante en el tiempo, se utiliza cuando la variable dependiente esté relacionada con la supervivencia de los individuos y se desee averiguar simultáneamente el efecto independiente una serie de factores sobre esta supervivencia.

Se debe tener en cuenta que no se trata sólo de saber el efecto sobre la supervivencia después de un tiempo determinado de seguimiento (por ejemplo, la supervivencia a los 5 años), sino de valorar cuál es el efecto sobre la función de supervivencia a lo largo de todo el periodo de observación de los pacientes, sea cual sea el punto temporal que se elija para la comparación. Si sólo interesase estudiar el efecto sobre la supervivencia en un punto del tiempo (por ejemplo, a los 5 años), entonces bastaría con un análisis de regresión logística, porque la variable respuesta sería dicotómica (sí sobreviven o no sobreviven).

En general, el Modelo de Regresión de Cox no hace ninguna suposición sobre la distribución de la variable aleatoria tiempo de supervivencia T, lo que implica que tampoco se hace ninguna suposición sobre la forma de $h_0(t)$, de ahí que sea considerado un modelo semiparamétrico.

Si suponemos que la función de riesgo de referencia tiene una distribución determinada estaremos ante un modelo paramétrico. Dependiendo de la suposición sobre la distribución de la función tendremos unos modelos u otros. Los modelos más comunes son los siguientes:

Modelo de riesgos proporcionales Weibull

$$h_0(t) = \lambda\,\gamma(t)^{\gamma-1} \;\Rightarrow\; h_i(t) = \lambda\,\gamma(t)^{\gamma-1}\,exp\{\beta' X_i(t)\} \qquad (4.12)$$

siendo $\lambda, \gamma > 0$ parámetros de escala y forma respectivamente. Si $\lambda > 1$ la tasa de riesgo aumenta, mientras que si $\lambda < 1$ la tasa de riesgo disminuye. En el caso particular de $\lambda = 1$ la tasa de riesgo permanece constante, y tendremos el modelo exponencial.

Modelo de riesgos proporcionales exponencial

$$h_0(t) = \lambda \quad \Rightarrow \quad h_i(t) = \lambda \, exp\{\beta' X_i(t)\} \qquad (4.13)$$

Modelo de riesgos proporcionales Gompertz

$$h_0(t) = \lambda \, exp\{\theta t\} \quad \Rightarrow \quad h_i(t) = \lambda \; exp\{\beta' X_i(t)\} \, exp\{\theta t\} \qquad (4.14)$$

El Modelo de Riesgos Proporcionales de Cox ha sido ampliamente utilizado en la estimación del riesgo cardiovascular.

Este modelo resolvía algunos de los problemas que presentaba la regresión logística. Las estimaciones mediante este tipo de regresión sólo permitían realizarse para cortos periodos de tiempo e ignoraban el tiempo hasta que ocurría el evento. Mediante el Modelo de Riesgos de Cox se resuelven estos dos aspectos.

El problema es que parte de un supuesto, el de la proporcionalidad, que exige que el riesgo relativo a diferencia del riesgo propiamente dicho no dependa del tiempo, lo que dicho de otra manera significa que sea constante a lo largo del tiempo. La literatura relacionada con el riesgo cardiovascular avala la idea de que el considerar efectos constantes a lo largo del tiempo para la variable edad y otros factores de riesgo del modelo(27) es una limitación que debe ser resuelta mediante el planteamiento de modelos que no partan de este supuesto(1) (34).

4.4. ACCELERATED FAILURE TIME REGRESSION MODEL

El Modelo de Riesgos Proporcionales de Cox descrito en el apartado 4.3 es el que comúnmente se ha utilizado para la estimación del riesgo en el campo de la medicina. Esta situación probablemente ha sido debida a que este modelo permite realizar estimaciones e inferencias sobre los parámetros sin suponer una distribución sobre el tiempo de supervivencia. El principal inconveniente de este modelo es la suposición de riesgos proporcionales a lo largo del tiempo. Esta suposición debe verificarse previamente a la aplicación del modelo para que las estimaciones que se obtengan sean

fiables. La realidad es que ese paso fundamental se omite y se presupone la proporcionalidad sin ninguna comprobación previa al respecto.

El *Accelerated Failure Time Regression Model*(12) *(AFT)* considera el tiempo hasta que ocurre el evento como la variable respuesta; lo que le permite la estimación de la probabilidad de presentar el evento para distintos intervalos de tiempo. Este modelo es un modelo paramétrico puesto que supone que la función de riesgo basal sigue una distribución determinada y mide el efecto de las variables explicativas directamente sobre la función de supervivencia, en vez de sobre la función de riesgo como hace el Modelo de Riesgos Proporcionales de Cox. La distribución que se supone en los modelos de riesgo cardiovascular es la distribución de Weibull, pero en este modelo se podrían emplear otras como la exponencial, log – logistic, log – Normal y gamma.

Sean,

X_{ij} el valor de la variable $j-ésima$ del individuo $i-ésimo$, siendo $i = 1, \dots, N$ y $j = 1, \dots, p$.

X_i el vector asociado al individuo $i-ésimo$, siendo $i = 1, \dots, N$.

En el modelo general de *Accelerated Failure Time Model* se define la función de riesgo para el individuo $i-ésimo$ como:

$$h_i(t) = e^{-\boldsymbol{\eta_i}} h_0\left(\frac{t}{e^{\boldsymbol{\eta_i}}}\right) \qquad (4.15)$$

Siendo $\boldsymbol{\eta_i} = \beta_1 x_{1i} + \ \dots + \beta_p x_{pi}$ la componente lineal del modelo, y $\boldsymbol{h_0(t)}$ la función de riesgo basal, al igual que se consideraba en el modelo de Cox.

La correspondiente función de supervivencia tiene la siguiente expresión:

$$S_i(t) = S_0\left(\frac{t}{\exp\left(\eta_i\right)}\right)$$

Considerando la forma Log – lineal del modelo para la variable tiempo T_i, asociada al tiempo de supervivencia del individuo $i-ésimo$, se tiene:

$$logT_i = \mu + \beta_1 x_{1i} + \ \dots + \beta_p x_{pi} + \sigma\varepsilon_i \qquad siendo \quad 1 \le i \le N \qquad (4.16)$$

siendo ε_i independientes e idénticamente distribuidas. Para cada distribución de la variable ε_i, obtendremos el correspondiente modelo para T. Por tanto según sea la distribución considerada, tenemos los siguientes modelos: *AFT Exponencial, AFT Weibull, AFT log-logístico, AFT log-normal y AFT gamma*.

En este modelo, los parámetros $\beta_j \ \ con\ 1 \leq j \leq p$ reflejan el efecto que cada variable explicativa tiene sobre el tiempo de supervivencia; valores positivos sugieren que el tiempo de supervivencia aumenta al aumentar los valores de las variables explicativas y viceversa.

La formulación Log – lineal del modelo puede ser utilizada para dar la expresión general de la función de supervivencia asociada a T_i que puede ser expresada a partir de la asociada a ε_i:

$$S_i(t) = P(T_i \geq t) = P(logT_i \geq logt)$$

Por tanto, a partir de (4.16) se tiene:

$$S_i(t) = P\big(\mu + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi} + \sigma\varepsilon_i \geq logt\big) =$$
$$= P\left(\varepsilon_i \geq \frac{logt - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}\right) \qquad \text{(4.17)}$$

Si lo expresamos en función de la función de supervivencia para la variable aleatoria ε_i, es decir en función de $S_{\varepsilon_i}(\varepsilon)$ se tiene que:

$$S_i(t) = S_{\varepsilon_i}\left(\frac{logt - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}\right) \qquad \text{(4.18)}$$

Este resultado muestra cómo la función de supervivencia para T_i puede ser expresada a partir de la función de supervivencia de la distribución de ε_i. Este resultado demuestra que el *Accelerated Failure Time Model* se puede obtener a partir de las distintas distribuciones de probabilidad consideradas para la variable ε_i.

Una expresión general del percentil $p - ésimo$ de la distribución de los tiempos de supervivencia se deriva de las relaciones anteriores. Sea $t_i(p)$ el percentil $p - ésimo$ para el individuo $i - ésimo$, entonces:

$$S_i\{t_i(p)\} = \frac{100 - p}{100}$$

A partir de la ecuación anterior, se tiene:

$$P\left(\varepsilon_i \geq \frac{logt_i(p) - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}\right) = \frac{100 - p}{100}$$

Si denotamos por $\varepsilon_i(p)$ el percentil $p - ésimo$ de la distribución de ε_i, entonces:

$$S_{\varepsilon_i}\{\varepsilon_i(p)\} = P\big(\varepsilon_i \geq \varepsilon_i(p)\big) = \frac{100 - p}{100}$$

Por tanto, necesariamente se tiene que:

$$\varepsilon_i(p) = \frac{logt_i(p) - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}$$

De donde,

$$t_i(p) = exp\big\{\sigma\varepsilon_i(p) + \mu + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi}\big\} \qquad (4.19)$$

es el percentil de la distribución de tiempos de supervivencia para el individuo $i - ésimo$. El percentil de la ecuación (4.19) puede ser expresado de manera análoga como sigue:

$$t_i(p) = exp\big(\beta_1 x_{1i} + \ \ldots + \beta_p x_{pi}\big)t_0(p)$$

Siendo $t_0(p)$ el percentil $p - ésimo$ asociado al individuo de referencia en el que todas las variables explicativas toman el valor cero.

Este hecho confirma que los coeficientes β_j $(1 \leq j \leq p)$ se interpretan en términos de efectos de las variables explicativas sobre un percentil determinado de la distribución de los tiempos de supervivencia.

La función de riesgo acumulada de la distribución de la variable T_i que denotamos por $H_i(t) = -logS_i(t)$, se tiene que:

$$H_i(t) = -logS_{\varepsilon_i}\left(\frac{logt - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}\right) =$$
$$= H_{\varepsilon_i}\left(\frac{logt - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}\right) \qquad (4.20)$$

donde $H_{\varepsilon_i}(\varepsilon_i) = -logS_{\varepsilon_i}(\varepsilon)$ es la función de riesgo acumulada para ε_i. La correspondiente función de riesgo, que se obtiene derivando la expresión $H_i(t)$ respecto de t en la ecuación (4.20), es:

$$h_i(t) = \frac{1}{\sigma t} h_{\varepsilon_i}\left(\frac{logt - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}\right) \qquad (4.21)$$

donde $h_{\varepsilon_i}(\varepsilon)$ es la función de riesgo para la distribución de ε_i.

Las distribuciones de ε_i más empleadas en este tipo de modelos son las mencionadas anteriormente y la razón principal es porque los percentiles $\varepsilon_i(p)$, tienen una forma simple.

4.4.1. Weibull Accelerated Failure Time Model

Un caso particular del *AFT model* empleado en el cálculo del riesgo cardiovascular, es ***Weibull AFT model***, en el que se supone que la variable aleatoria tiempo $T \equiv W(\lambda, \gamma)$, siendo $\lambda, \gamma > 0$ parámetro de escala y forma respectivamente. En este caso la función de riesgo basal se expresa:

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

La función de riesgo para el individuo $i-$ésimo a partir de la expresión (4.15) viene dada por:

$$h_i(t) = e^{-\eta_i}\lambda\gamma(e^{-\eta_i}t)^{\gamma-1} = (e^{-\eta_i})^{\gamma}\lambda\gamma t^{\gamma-1}$$

por tanto, el tiempo de supervivencia del individuo $i-$ésimo sigue una distribución $W(\lambda e^{-\gamma\eta_i}, \gamma)$, por lo que bajo diferentes valores de las covariables difiere sólo en el parámetro de escala.

La distribución de Weibull se dice que posee la propiedad AFT *("accelerated failure time").* De hecho, esta es la única distribución de probabilidad que verifica la propiedad de riesgos proporcionales y *AFT*.

Debido a que la distribución de Weibull tiene ambas propiedades, riesgos proporcionales y *AFT*, hay una correspondencia directa entre los parámetros de los dos modelos.

Si la función de riesgo basal sigue la distribución $W(\lambda,\ \gamma)$, el tiempo de supervivencia según el modelo de riesgos proporcionales sigue una distribución $W(\lambda exp(\alpha' x_i),\ \gamma)$, mientras que según el modelo *AFT* sigue la distribución $W(\lambda exp(-\gamma\beta' x_i),\ \gamma)$. Por tanto, si multiplicamos los coeficientes de las variables explicativas de la componente lineal del modelo AFT por $-\gamma$, se obtienen los correspondientes $\alpha - coeficientes$ del modelo de riesgos proporcionales.

En términos de la representación log – lineal del modelo en la ecuación (4.16) si T_i sigue una distribución Weibull, entonces ε_i tiene en cierto modo la distribución conocida como *distribución de Gumbel*. Esta distribución es asimétrica y su función de supervivencia viene dada por:

$$S_{\varepsilon_i}(\varepsilon) = exp\big(-exp\ (\varepsilon)\big)$$

con $-\infty < \varepsilon < \infty$. La función de riesgo acumulado y función de riesgo de esta distribución vienen dadas por $H_{\varepsilon_i}(\varepsilon) = e^{\varepsilon}$ y $h_{\varepsilon_i}(\varepsilon) = e^{\varepsilon}$, respectivamente.

Para mostrar que la variable aleatoria $T_i = exp(\mu + \boldsymbol{\beta' x_i} + \sigma\varepsilon_i)$ tiene una distribución de Weibull, a partir de la expresión (4.18) la función de supervivencia de T_i viene dada por:

$$S_i(t) = exp\left\{-exp\left(\frac{logt - \mu - \beta_1 x_{1i} - \ \ldots - \beta_p x_{pi}}{\sigma}\right)\right\} \qquad \textbf{(4.22)}$$

Esto puede ser expresado de la forma:

$$S_i(t) = exp\big(-\lambda_i t^{1/\sigma}\big)$$

donde

$$\lambda_i = exp\left\{-\frac{\big(\mu + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi}\big)}{\sigma}\right\}$$

y que por tanto coincide con la función de supervivencia de la distribución Weibull con parámetro de escala λ_i y parámetro de forma σ^{-1}. Por ello, la ecuación (4.22) es la representación del modelo AFT de la función de supervivencia del modelo de riesgos proporcionales de Weibull descrito en el apartado anterior.

La función de riesgo acumulado y la función de riesgo para el modelo *AFT Weibull* se pueden obtener bien directamente a partir de la ecuación (4.22) o bien a partir de $H_{\varepsilon_i}(\varepsilon)$ y $h_{\varepsilon_i}(\varepsilon)$ utilizando los resultados obtenidos en (4.20) y (4.21).

La función de riesgo acumulada se expresa por tanto:

$$H_i(t) = -logS_i(t) = exp\left(\frac{logt - \mu - \beta_1 x_{1i} - \ ... - \beta_p x_{pi}}{\sigma}\right)$$

que puede también ser expresada como $\lambda_i t^{1/\sigma}$, y la función de riesgo viene dada por:

$$h_i(t) = \frac{1}{\sigma t} exp\left(\frac{logt - \mu - \beta_1 x_{1i} - \ ... - \beta_p x_{pi}}{\sigma}\right) \qquad (4.23)$$

o análogamente $h_i(t) = \lambda_i \sigma^{-1} t^{\sigma^{-1}-1}$.

A partir de la expresión anterior del modelo se comprueba la coincidencia con el modelo de riesgos proporcionales de Weibull. La función de supervivencia por tanto para el individuo $i - ésimo$ es:

$$S_i(t) = exp\{-exp(\alpha_1 x_{1i} + \ ... + \alpha_p x_{pi})\lambda t^{\gamma}\} \qquad (4.24)$$

siendo λ y γ los parámetros de la función de riesgo basal de Weibull. Por tanto, hay una correspondencia directa entre la ecuación (4.22) y la ecuación (4.24) en el sentido siguiente:

$$\lambda = exp\left(-\frac{\mu}{\sigma}\right) \qquad \gamma = \sigma^{-1} \qquad \alpha_j = -\frac{\beta_j}{\sigma}$$

para $j = 1, ..., p$. Entonces se deduce que el modelo log – lineal se expresa:

$$logT_i = \frac{1}{\gamma}\{-log\lambda - \alpha_1 x_{1i} - \ ... - \alpha_p x_{pi} + \varepsilon_i\}$$

y donde ε_i sigue una distribución de Gumbel, lo que nos muestra una representación alternativa del modelo de riesgos proporcionales de Weibull.

Según esta forma del modelo, el percentil $p - ésimo$ para el tiempo de supervivencia del individuo $i - ésimo$ es el valor $t_i(p)$, que es según la ecuación (4.22):

$$t_i(p) = exp\left[\sigma log\left\{-log\left(\frac{100 - p}{100}\right)\right\} + \mu + \boldsymbol{\beta}'\boldsymbol{x_i}\right] \qquad (4.25)$$

De manera análoga, el percentil $p-ésimo$ según la distribución de ε_i, $\varepsilon_i(p)$ será:

$$exp\{-e^{\varepsilon_i(p)}\} = \frac{100-p}{100},$$

de donde,

$$\varepsilon_i(p) = log\left\{-log\left(\frac{100-p}{100}\right)\right\}$$

Y el resultado general de la ecuación (4.19) nos lleva directamente a la expresión (4.25).

La función de supervivencia y la función de riesgo del modelo de Weibull que se derivan de las ecuaciones (4.22), (4.23) y de la ecuación (4.25) permiten que los percentiles sean estimados directamente.

4.4.2. Ajuste del *Accelerated Failure Time Regression Model*

El ajuste del modelo AFT se realiza mediante el método de máxima verosimilitud.

La función de verosimilitud considerados los tiempos de supervivencia $t_1, \dots, t_N$:

$$L(\beta, \mu, \sigma) = \prod_{i=1}^{N} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i}$$

Siendo $f_i(t_i)$ y $S_i(t_i)$, las funciones de densidad y supervivencia respectivamente del individuo $i-ésimo$ en el instante t_i; δ_i el indicador de si se produce o no fallo en el individuo $i-ésimo$.

A partir de la relación $S_i(t_i) = S_{\varepsilon_i}(z_i)$, donde $z_i = \frac{(logt_i - \mu - \beta_1 x_{1i} - \dots - \beta_p x_{pi})}{\sigma}$, derivando respecto de t, se tiene:

$$f_i(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}(z_i)$$

Por ello, la función de verosimilitud expresada en términos de la función de supervivencia y de densidad de ε_i se expresa:

$$L(\beta,\mu,\sigma)=\prod_{i=1}^{N}(\sigma t_i)^{-\delta_i}\{f_{\varepsilon_i}(z_i)\}^{\delta_i}\{S_{\varepsilon_i}(z_i)\}^{1-\delta_i}$$

Considerando logaritmos tenemos:

$$logL(\beta,\mu,\sigma)=\sum_{i=1}^{N}\{-\delta_i\, log(\sigma t_i)+\delta_i log f_{\varepsilon_i}(z_i)+(1-\delta_i)logS_{\varepsilon_i}(z_i)\}$$

Por tanto, se tienen $p+2$ parámetros desconocidos, $\mu,\ \sigma,\ \beta_1,\ \ldots,\beta_p,$ que deben ser estimados maximizando la función anterior. El método empleado para su resolución es el método iterativo de Newton – Raphson.

5. LIMITACIONES DE LOS MODELOS ACTUALES DE RIESGO CARDIOVASCULAR

5.1. INTRODUCCIÓN

El desarrollo de modelos matemáticos para la estimación del riesgo cardiovascular ha supuesto un gran avance para el conocimiento, interpretación y la modificación de los factores de riesgo relacionados con la enfermedad cardiovascular. Sin embargo, esta aproximación al riesgo tiene sus limitaciones, tanto por la metodología como por el enfoque planteado.

A continuación describimos las limitaciones fundamentales encontradas en los modelos de estimación estudiados. Aunque es difícil caracterizar la tipología de todas ellas atendiendo a una sola característica, distinguiremos los siguientes tipos con el fin de destacar el que consideramos es el origen de las limitaciones de los sistemas de predicción del riesgo cardiovascular planteados:

- Limitaciones metodológicas esenciales, considerando en este apartado aquellas que son inherentes a la propia metodología estadística empleada.
- Otras limitaciones. En este apartado estudiaremos las limitaciones que podríamos considerar de origen epidemiológico, clínico y otras con una clasificación no tan clara pero que en definitiva son limitaciones encontradas y avaladas por la literatura.

Destacar en este punto que un planteamiento erróneo del modelo de estimación por la tipología de los datos o por la falta de comprobación de los supuestos del mismo, puede llevarnos a realizar estimaciones incorrectas y a observar limitaciones que podrían ser consideradas clínicas pero que en realidad el origen de las mismas sea una mala especificación del modelo de estimación del riesgo.

5.2. LIMITACIONES METODOLÓGICAS

La existencia de múltiples tablas de estimación del riesgo cardiovascular ha generado cierta confusión entre los profesionales, al no existir un acuerdo unánime respecto a qué sistema se debe emplear en España. Las tablas que posiblemente tienen más aceptación en España son las de *Framingham REGICOR* y *SCORE*, aunque el mayor consenso en España respecto a escalas cuantitativas se ha obtenido con la escala SCORE para países de bajo riesgo(35).

La concordancia entre distintas escalas es pobre(36) y la sensibilidad y especificidad de las escalas no son óptimas, lo que genera desconfianza en cuanto a su utilidad clínica. Buitrago Ramírez et al.(37) han publicado un estudio retrospectivo longitudinal de seguimiento a 10 años de 608 pacientes atendidos en una consulta de Atención Primaria comparando el rendimiento diagnóstico de los sistemas SCORE y de REGICOR. Los valores de sensibilidad y especificidad para SCORE fueron 66,7% y 91,7% respectivamente, y para REGICOR fueron 12,3% y 92,6%. Esta falta de concordancia y los valores no óptimos de sensibilidad y especificidad han sido motivos, entre otros, de búsqueda de alternativas al riesgo cardiovascular.

5.2.1. Limitaciones del Estudio de Framingham

Una de las limitaciones más importantes de este estudio es que al estar basado en una población de riesgo basal alto, sobrevalora el riesgo de poblaciones como la mediterránea. Por tanto, las predicciones que a partir de estas tablas se hagan para predecir el riesgo cardiovascular en nuestro país deben ser aceptadas con cautela, ya que claramente sobrestiman el riesgo en las poblaciones con tasas de incidencia de enfermedad coronaria bajas(38).

Según se ha descrito en capítulos anteriores, esta conocida limitación ha tratado de ser resuelta mediante la calibración de las tablas, es decir adaptándolas a la realidad epidemiológica del país para su correcta aplicación. En España, dada la ausencia de estudios poblacionales de cohorte, la calibración comúnmente empleada es la que se ha realizado a través del registro poblacional de infarto de miocardio de Girona REGICOR (Registre Gironí del Cor) del que se obtuvo la tasa de incidencia de acontecimientos mayores. Dado que la tasa de incidencia de angina y de IAM silente era desconocida en Girona, se asumió una proporción similar a la de Framingham.

Las tablas que a partir de esta calibración se obtuvieron, no han sido validadas por el procedimiento de base poblacional como el realizado para los acontecimientos mayores, ni por un procedimiento prospectivo. Si unimos este característica al hecho de que la incidencia de IAM en Girona se encuentra aproximadamente un 15% por debajo del promedio de España en el estudio IBERICA (Investigación y Búsqueda Específica y

Registro de Isquemia Coronaria Aguda), concluimos que la validez externa de la ecuación a otras zonas de España debe aceptarse con las debidas precauciones(21).
En cuanto a la metodología estadística empleada, según se ha descrito en 2.1 MODELOS DE FRAMINGHAM, las tablas de riesgo se elaboran partiendo de dos métodos estadísticos: el modelo de Wilson (1998) (13), el de Framingham D'Agostino (2008) (17) y Framingham a 30 años (2009) (18) se basan en el **Modelo de Riesgos Proporcionales de Cox**(14); mientras que el modelo clásico de Anderson (1991) (9) y el de Framingham D'Agostino (2000) (16) emplean el modelo **Accelerated Failure Time**(12) suponiendo que la variable aleatoria tiempo de supervivencia se distribuye según una distribución de Weibull.
El Modelo de Riesgos Proporcionales de Cox(14) ha sido uno de los modelos de regresión lineal más populares en el análisis de la supervivencia durante las últimas décadas, especialmente en la literatura médica. Dada la importancia en la toma de decisiones que se pueden alcanzar a partir de estos análisis, la falta de robustez o una mala especificación del modelo son cuestiones importantes a tener en cuenta. Una mala especificación o falta de ajuste puede darse por existir covariables omitidas, por la omisión del tratamiento en la interacción de las variables, o la violación de los supuestos del modelo entre otros.

En el caso del Modelo de Riesgos Proporcionales de Cox, se ha de verificar el supuesto de proporcionalidad, que supone que el riesgo relativo a diferencia del riesgo propiamente dicho no depende del tiempo; o dicho de otra manera, que sea constante a lo largo del tiempo.

La literatura relacionada con el riesgo cardiovascular avala la idea de que el considerar efectos constantes a lo largo del tiempo para la variable edad y otros factores de riesgo del modelo es una limitación(27) que debe ser resuelta mediante el planteamiento de modelos que no partan de este supuesto(1) (34).

Para valorar la calidad del ajuste del modelo de Cox hay test que están basados en ideas similares a las de Hosmer y Lemeshow(39), pero su éxito para detectar una especificación errónea no está garantizado. Incluso con valores de bondad de ajuste buenos puede que el modelo no ajuste bien y este hecho sea debido a que la suposición de proporcionalidad no ha sido verificada previamente y por tanto el modelo no sea el correcto desde el planteamiento del estudio(40).

Se podría plantear el calibrar las ecuación de Framingham a partir de alguna de las otras versiones que están basadas en el modelo **Accelerated Failure Time**(12) como el modelo clásico de Anderson (1991) y el de Framingham D'Agostino (2000), en el que se supondría que la distribución del tiempo de supervivencia es conocida. Sin embargo, la suposición del modelo de que la distribución de la supervivencia bajo diferentes valores de las covariables difiere sólo en la escala es también una suposición bastante fuerte(41). Por ello, se hace necesaria la propuesta de modelos más sofisticados que resuelvan las limitaciones destacadas.

5.2.2. Limitaciones de European Heart SCORE

La función de riesgo *EuroSCORE* estima el riesgo mortal de todas las manifestaciones aterotrombóticas cardiovasculares, incluidos el ictus, la insuficiencia cardiaca, la insuficiencia arterial periférica o ciertos aneurismas y no sólo la enfermedad coronaria. Un aspecto controvertido de estas tablas es el hecho de que no tienen en cuenta los acontecimientos no letales, lo que puede alejarlas muchas veces del objetivo del cálculo de riesgo en prevención primaria en la práctica clínica, que es identificar a los pacientes no sólo con mayor riesgo de morir, sino también de presentar un episodio cardiovascular que pueda causar secuelas y afectar a su calidad de vida.

En cuanto a la metodología estadística empleada, según se ha descrito en el capítulo 3, el cálculo del riesgo con un horizonte temporal de 10 años se realiza mediante el **Modelo de Riesgos Proporcionales de Weibull**, es decir el Modelo de Riesgos Proporcionales de Cox(14) en el que se supone una distribución Weibull para la variable aleatoria tiempo de supervivencia.

En España las tablas más extendidas en cuanto a su uso en atención primaria son:

- *Framingham REGICOR*

  Las tablas empleadas en nuestro país de *Framingham REGICOR*(21), adaptan las ecuaciones de *Framingham Wilson*(13) (1998) a la realidad española a partir de los datos del registro poblacional de infartos de miocardio de Gerona REGICOR (Registre Gironí del Cor)(22) adaptando la prevalencia de los

factores de riesgo cardiovascular y la tasa de incidencia de eventos coronarios a los de nuestro medio. Dado que el modelo del que se derivan es el de *Framingham Wilson*(13), la estimación del riesgo se realiza mediante el **Modelo de Riesgos Proporcionales de Cox**(14).

- *European Heart SCORE*

  En este caso hay dos versiones de tablas de estimación del riesgo que pueden ser empleadas en nuestro país. Las tablas del estudio original para regiones de bajo riesgo o la versión *EuroSCORE* calibrada mediante la utilización de las tasas de mortalidad española y los factores de riesgo del estudio MONICA – Catalunya, cuyo planteamiento surgió ante las recomendaciones de las guías de prevención de adaptar el modelos *EuroSCORE* al nivel de riesgo de cada país. En cualquiera de los casos, el modelo de estimación del riesgo es el **Modelo de Riesgos Proporcionales de Weibull**, es decir el modelo de riesgos proporcionales de Cox(14) en el que se supone una distribución Weibull para la variable aleatoria tiempo de supervivencia.

Por tanto, la metodología empleada en *Framingham REGICOR* y en *EuroSCORE* se basa en el Modelo de Riesgos Proporcionales de Cox(14), y por tanto se ha de verificar la condición de proporcionalidad que requiere el mismo, es decir, que el riesgo relativo a diferencia del riesgo propiamente dicho no depende del tiempo.

El mayor consenso en España(35) respecto a escalas cuantitativas se ha obtenido con la escala *EuroSCORE* para países de bajo riesgo, aunque no está exenta de limitaciones. Este consenso se ha obtenido a pesar de que son necesarios estudios de cohorte en nuestro medio para validar las tablas *EuroSCORE*(3) mientras que la función REGICOR ya está validada(42).

En nuestro país recientemente se ha publicado un modelo de estimación denominado ERICE – Score(5) en el que se propone una nueva ecuación de riesgo cardiovascular genuinamente española obtenida a partir del riesgo concurrente individual de los participantes de siete cohortes españolas de población de mediana edad y anciana. La ecuación ERICE ofrece una estimación del riesgo cardiovascular total a diez años,

teniendo en cuenta los factores de riesgo habituales en otras ecuaciones y otros habitualmente no incluidos como la diabetes mellitus y el tratamiento farmacológico de los factores de riesgo cardiovascular. La predicción del riesgo se realiza a través del modelo de riesgos proporcionales de Cox, mediante el que se examina la contribución de los diferentes factores considerados como independientes, al riesgo de cualquier evento cardiovascular (mortal y no mortal).

Por tanto, esta ecuación da respuesta a la necesidad de realizar las estimaciones basándose en estudios de cohorte del país de origen donde se pretende realizar la estimación, pero el modelo al igual que en el resto de sistemas de cuantificación descritos parte del supuesto de proporcionalidad, lo que supone en sí misma una limitación.

De lo anterior se puede concluir que la limitación fundamental de las distintas tablas utilizadas en prevención primaria en España se debe en la metodología estadística seleccionada, el Modelo de Riesgos Proporcionales de Cox(14).

Para que las predicciones de los estudios de *Framingham REGICOR* y *SCORE* puedan considerarse fiables se debería verificar previamente el supuesto de proporcionalidad inherente al modelo de Cox(14), es decir, que el riesgo relativo a diferencia del riesgo propiamente dicho no depende del tiempo. Esta suposición se hace a pesar de que la literatura avala la idea de que el considerar efectos constantes a lo largo del tiempo para la variable edad y otros factores de riesgo del modelo es una limitación(27) que debe ser resuelta mediante el planteamiento de modelos que no partan de este supuesto(34) (1).

Otra de las limitaciones a destacar que comparten los sistemas de cuantificación del riesgo cardiovascular descritos, es referente a los factores de riesgo considerados y a la metodología empleada para su selección.

- Las tablas de riesgo cardiovascular actuales estiman el riesgo a partir de un número reducido de variables y prácticamente el mismo conjunto de factores de riesgo en todas ellas. En el caso de las ecuaciones de *Framingham por catetorías de Wilson*, modelo a partir del cual se obtienen las tablas calibradas de *Framingham – REGICOR* y *Framingham – DORICA*, incluyen como factores de riesgo la edad, sexo, presión arterial sistólica y diastólica, colesterol HDL,

colesterol LDL y las variables diabetes mellitus y tabaquismo. En el caso de *SCORE* se estratifica por cohorte y sexo, estimando el modelo mediante la edad, colesterol LDL, colesterol HDL, niveles de triglicéridos, antecedentes familiares de infarto agudo de miocardio, presión arterial sistólica, diabetes y tabaquismo. Aunque se considera que los factores de riesgo están bien identificados(43), en la actualidad se plantea la búsqueda de otros factores de riesgo(3) (27).

- La metodología diseñada para la selección de las variables predictoras tiende a seleccionar solamente aquellas variables que presentan una cierta correlación lineal con el evento, ya que para su selección se ha variado sucesivamente un factor manteniendo el resto constantes. Esta aproximación tiende a seleccionar aquellas variables que presentan una cierta correlación lineal con el evento, resultando claramente limitada para aproximarse a la complejidad de las relaciones fuertemente no lineales entre los factores implicados en el riesgo.

A pesar de que el desarrollo de modelos matemáticos para la estimación del riesgo cardiovascular ha supuesto un gran avance para el conocimiento, interpretación y lo que es más importante, la modificación de los factores de riesgo relacionados con la enfermedad cardiovascular, es cierto que la aproximación al riesgo realizada no deja de tener limitaciones esenciales e inherentes a la propia metodología que comprometen el resultado final desde su propio diseño. El planteamiento erróneo de un modelo de estimación puede llevarnos a realizar estimaciones incorrectas y a observar limitaciones que podrían ser consideradas clínicas pero que en realidad el origen de las mismas sea una mala especificación del modelo de estimación del riesgo.

5.3. OTRAS LIMITACIONES

En este apartado estudiaremos las limitaciones que consideramos de origen epidemiológico, clínico y otras con una clasificación no tan clara pero que en definitiva son limitaciones encontradas y avaladas por la bibliografía.

- El riesgo estimado está diseñado para una población de origen con características específicas y presenta una fuerte dependencia de la misma. Por

ello, los resultados no deben extrapolarse directamente a poblaciones que presenten diferentes niveles de riesgo cardiovascular. Esta limitación según se ha comentado en los apartados 2.2 y 3, ha tratado de ser solventada mediante la calibración de las tablas originales. A pesar de ello se sigue insistiendo en la necesidad de obtener ecuaciones a partir de estudios de cohortes del país de origen.

- El tratamiento del riesgo como un proceso estático en el tiempo. Las funciones de riesgo estiman el riesgo de la aparición de un evento CVD con un horizonte temporal fijo (10 años en general), con lo que el riesgo se reduce a una probabilidad, perdiéndose la perspectiva dinámica de la evolución del riesgo en el tiempo. Sin especificar la enorme diferencia que existe entre que el episodio ocurra al mes siguiente del cálculo del riesgo y que lo haga nueve años y once meses después(44). Y no sólo eso, sino que además las estimaciones con ese horizonte temporal prefijado que habitualmente es de diez años, se realizan a través de las mediciones de los factores de riesgo en el momento actual. Este hecho resta precisión y fiabilidad a las estimaciones realizadas.

- Suponen que el riesgo asociado a los FR incluidos en ellas se mantiene constante toda la vida(34), cuando sabemos que en algunos casos no es así. Por ejemplo, es conocido que a partir de los 65 años la fracción lipídica que predice mejor el RCV es el cHDL, y en la ecuación de Framingham se constató que el colesterol(45) total sólo tenía valor predictivo en los menores de 50 años(27).

- Las funciones no tienen en cuenta el tiempo de exposición a los diferentes factores de riesgo considerados y la mayoría no tienen en cuenta los tratamientos farmacológicos (3) (46).

- Falta de concordancia entre las tablas de riesgo utilizadas en prevención primaria en nuestro país. En relación a esta falta de concordancia destacar los siguientes estudios en los que se avala la falta de concordancia planteada:

"Comparación entre la tabla del SCORE y la tabla Framingham REGICOR en la estimación del riesgo cardiovascular en una población urbana seguida durante 10 años"(37).

Este estudio afirma que ninguna de las tablas, *SCORE* y *Framingham REGICOR*, obtiene unos criterios de validez óptimos. La función de *REGICOR* obtuvo una sensibilidad del 12,3%, lo que significa que tendría un porcentaje del 87,7% de falsos negativos, es decir, de pacientes a los que *REGICOR* catalogaría como de riesgo cardiovascular no alto sin serlo, y a quienes, por lo tanto, se podría estar privando del beneficio de fármacos antihipertensivos y/o hipolipemiantes. La sensibilidad es menor en mujeres, en quienes el porcentaje de falsos negativos llegaría al 90% en *REGICOR* frente al 66,7% en *SCORE*. En varones, *SCORE* consigue unos indicadores de sensibilidad y especificidad más igualados (del 83,3% y el 84,0%, respectivamente) frente al 13,5% y el 85% de *REGICOR*. Por lo tanto, la posibilidad de que un paciente varón dé como resultado un falso negativo se quintuplica en REGICOR frente al SCORE. En varones de 50 – 65 años *SCORE* logra una sensibilidad del 100% y una especificidad del 70%, frente al 14,8% y el 70,4%, respectivamente con *REGICOR*, lo que indicaría que en ese grupo de edad el SCORE no generaría falsos negativos y tendría un 30% de falsos positivos, es decir, pacientes a los que se consideraría de riesgo alto y, por tanto, susceptibles de tratamiento farmacológico, sin en realidad serlo. En el grupo de las mujeres mayores de la cohorte (60 – 65 años) se obtuvieron los mejores indicadores SCORE, con una sensibilidad del 100% y una especificidad del 92,6%.

La comparación entre el riesgo global estimado en la cohorte por las funciones *REGICOR* y *SCORE* y el porcentaje real de episodios coronarios y muertes cardiovasculares ocurridos en la población revela que *REGICOR* subestimó el riesgo cardiovascular (un 4,9% frente al 7,9%; $p<0,001$), mientras que *SCORE* sobrestimó el riesgo de muerte cardiovascular (un 2,1% frente al 1,5%; $p<0,001$).

"Impacto de la utilización de las diferentes tablas SCORE en el cálculo de riesgo cardiovascular"(46).

Este estudio alude a la falta de concordancia en las distintas versiones disponibles de SCORE. La función de riesgo *SCORE* calibrada identifica a más

pacientes de alto riesgo que *SCORE* para países de bajo riesgo y *SCORE* con colesterol unido a lipoproteínas de alta densidad, por lo que su utilización implicaría tratar a más pacientes con estatinas.

"Riesgo cardiovascular del SCORE comparado con el de Framingham. Consecuencias del cambio propuesto por las Sociedades Europeas"(47).

Los resultados de este estudio plantean que el cambio propuesto por las Sociedades Europeas tendrá repercusiones cuantitativas y cualitativas, y en algunas ocasiones contrarias a las pruebas disponibles de los ensayos clínicos con fármacos hipolipemiantes. El porcentaje y las características de los pacientes de riesgo alto son diferentes. Las Sociedades Europeas introducen un nuevo paciente de riesgo alto, numéricamente poco importante, caracterizado por ser mujer en la mayoría de los casos, edad avanzada, presión arterial elevada y colesterol normal. Aquí, las pruebas científicas de la eficacia del tratamiento con fármacos hipolipemiantes son escasas. Por otro lado, dejarán de considerar de riesgo alto y candidato al tratamiento hipolipemiante a un grupo de pacientes varones mayoritariamente, de 60 años de edad, colesterol más elevado y presión arterial más baja. Precisamente este es el grupo más numeroso y donde son más fuertes las evidencias de la eficacia del tratamiento con fármacos hipolipemiantes

- Otra limitación importante es su baja sensibilidad, ya que gran parte de los acontecimientos coronarios o cardiovasculares se presentan en el grupo de la población con riesgo intermedio(3). Esta aparente paradoja se explica atendiendo a que una gran proporción de la población tiene riesgo intermedio y por lo tanto aporta muchos casos.

  Se están haciendo esfuerzos importantes para identificar biomarcadores que mejoren la reclasificación de individuos sobre todo de riesgo intermedio, tales como los triglicéridos, la proteína C reactiva, características genéticas. Igualmente otros factores que actualmente tampoco figuran en las funciones de riesgo de la enfermedad CV, como antecedentes familiares de enfermedad CV, obesidad, pulso pedio, índice tobillo – brazo, grosor de la íntima – media carotídeo o la proteinuria, podrían contribuir a mejorar la predicción(48).

No obstante, hay que señalar que, aunque se han identificado numerosos factores de riesgo distintos de los incluidos en las funciones de riesgo disponibles, como concentraciones de proteína C reactiva y homocisteína, su contribución a la estimación del riesgo CV total para pacientes individuales (aparte de los factores tradicionales de riesgo) es generalmente baja(49). Por ello, la inclusión en el modelo de nuevos factores de riesgo podría ser útil para una correcta reclasificación de aquellos pacientes considerados con riesgo intermedio para priorizar intervenciones preventivas(45).

- La estimación del riesgo no es fiable en las franjas de edad jóvenes y edad avanzada:
    - Los individuos jóvenes siempre presentan un riesgo absoluto bajo, aunque el resto de factores de riesgo sean todos desfavorables(45). La estimación del riesgo en personas jóvenes con un riesgo absoluto bajo pero un riesgo relativo de enfermedad cardiovascular alto sigue siendo controvertido. Las funciones estiman el riesgo a 10 años, y en personas jóvenes este riesgo suele ser bajo, ya que la edad es el principal determinante del riesgo. En este sentido hay estudios que tratan de resolver esta problemática mediante el cálculo de la edad vascular(43), (50), la estimación del riesgo a 30 años y a lo largo de la vida(51), (48).

    - La estimación del riesgo en la franja de edad avanzada sigue siendo un reto. En algunas categorías de edad, la mayoría de las personas, especialmente los varones, tendrán una estimación de riesgo cardiovascular superior al 5 – 10% con base únicamente la edad (y el sexo), incluso cuando los niveles de otros factores de riesgo CV sean relativamente bajos. Esto puede llevar al uso excesivo de fármacos. Además ya que la mayor parte de las enfermedades cardiovasculares se van a producir en los mayores de 65 años o incluso en los mayores de 74 años, en muchos pacientes no podremos calcular el riesgo cardiovascular porque están fuera del intervalo de edad para el cual el modelo se ha diseñado. Por ejemplo, las tablas SCORE solamente se pueden aplicar a personas entre 40 y 65 años. Además parece razonable desarrollar

modelos que calculen el riesgo en esta franja de edad avanzada a medio plazo, como podrían ser 5 años(27).

El desarrollo de modelos matemáticos para la estimación del riesgo cardiovascular ha supuesto un gran avance para el conocimiento, interpretación y la modificación de los factores de riesgo relacionados con la enfermedad cardiovascular. Sin embargo, el estudio realizado pone de manifiesto una falta de concordancia entre los diferentes modelos y que existen una serie de limitaciones, inherentes o no a la propia metodología, que deben ser solventadas mediante nuevos planteamientos.

Si hay algo claro a la hora de plantear un sistema de predicción del riesgo cardiovascular es que los factores de riesgo cardiovascular no deben analizarse de forma separada, ya que el riesgo que presenta un individuo con más de un factor es superior a la suma de cada uno de ellos, es decir, no es aditivo. Por ello, no cabe ninguna duda que la estimación del riesgo debe hacerse desde una perspectiva multivariante. Además para que el modelo estadístico seleccionado sea fiable, se deben verificar previamente los supuestos de los que parte para su aplicación. En el caso de España, las tablas de estimación del riesgo utilizadas en prevención primaria parten del Modelo de Riesgos Proporcionales de Cox(14) que como ya se ha mencionado, considera efectos independientes de las variables explicativas y supone que el riesgo relativo no depende del tiempo. Esta suposición se hace a pesar de que la literatura avala la idea de que el considerar efectos constantes a lo largo del tiempo para la variable edad y otros factores de riesgo del modelo es una limitación(27). Este planteamiento compromete el resultado final desde su propio diseño.

Además, según un estudio reciente realizado por 1.390 médicos de atención primaria de España, se observa que únicamente el 38% de los profesionales calculaba el riesgo en más del 80% de sus pacientes con al menos un factor de riesgo cardiovascular (52). Las principales barreras para el cálculo del riesgo cardiovascular señaladas por los profesionales fueron la falta de tiempo (81%), la falta de calculadoras de riesgo informatizadas (19%), que las funciones no se basan en datos obtenidos en la población

española (16%) y la falta de información sobre alguna variable necesaria para el cálculo del riesgo (15%).

Por tanto, para conseguir una aplicación real del modelo diseñado éste debe tratar de solventar tanto estas barreras como las limitaciones descritas a lo largo del apartado 5.3.

Otro de los requerimientos que debe caracterizar al modelo es que sea fácilmente interpretable. Este requerimiento es clave en el planteamiento de nuevos modelos de estimación del riesgo cardiovascular. El principal motivo es que el modelo de predicción del riesgo se diseña para su uso en prevención primaria, donde el objetivo no es sólo estimar el riesgo del paciente sino que también es necesario conseguir la motivación del mismo en el cumplimiento terapeútico relacionado con los factores modificables. Por ello, el nuevo modelo de predicción debe permitir, además de valorar y explicar la aportación al riesgo cardiovascular individual de cada uno de los factores incluidos en el modelo, ser una herramienta útil para cuantificar la evolución del riesgo en función de la modificación de los distintos factores de riesgo que se presenten.

6. PROPUESTA DE MODELO DE ESTIMACIÓN DEL RIESGO CARDIOVASCULAR

6.1. INTRODUCCIÓN

Una de las piezas claves de este capítulo y motivación principal de este trabajo es la Tesis Doctoral *"Modelos Multivariantes Internos de Medición de Riesgos de Crédito, Acordes con Basilea II"*(53) de Fernando Mallo Fernández (2011). En ella se lleva a cabo la investigación sobre el desarrollo de mejores modelos proactivos de *credit scoring* desde la óptica de Basilea II. Una de las contribuciones de su trabajo es el de presentar una visión unificada de las técnicas de *credit scoring*, formalizando sus estructuras funcionales como expansión de funciones de base. Otra de sus aportaciones y que es la base de nuestra propuesta es la formulación de los Modelos Logísticos Lineales Híbridos que constituyen una extensión natural de los modelos logísticos lineales a los que se añade la capacidad para recoger la no linealidad de las variables explicativas.

El objetivo de este trabajo es el de trasladar al campo de la medicina las aportaciones del trabajo de F. Mallo Fernández en relación al planteamiento y estimación de los modelos de riesgo de crédito, con el objetivo de proponer un sistema de predicción del riesgo cardiovascular basado en Modelos Logísticos Lineales Híbridos.

En el caso del riesgo de crédito, el objetivo del planteamiento es el de estimar la probabilidad asociada a una variable respuesta binaria indicadora de la presencia o no de impago. Esta estimación será valorada a partir de una serie de variables explicativas que se denominan factores de riesgo. Este mismo planteamiento se trasladará al campo médico, en el que el objetivo es el de estimar la probabilidad asociada a una variable binaria indicadora en este caso de la presencia o no de enfermedad cardiovascular a partir de los factores de riesgo que sean seleccionados para formar parte del modelo.

Por tanto el símil entre ambos planteamientos es evidente y la aplicación con éxito del modelo planteado en el riesgo de crédito es previsible que lo sea para la estimación del riesgo cardiovascular que nos ocupa.

6.2. CONCEPTOS GENERALES

Sean $\{(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1,\dots,N}$ los datos observados sobre los individuos de la muestra considerada que en principio tiene una distribución $F_{X,Y}(x, y)$ desconocida; siendo X_i el

vector de variables explicativas del individuo i – ésimo y siendo Y_i la variable binaria indicadora del evento cardiovascular. Denotaremos por $f_{X,Y}(x,y)$ la función de densidad, cuya estimación será fundamental para conocer la relación entre el estado del riesgo y las características observadas en el individuo muestreado, así como para estimar su riesgo cardiovascular, clasificarlo atendiendo al nivel de riesgo y conocer la influencia que cada variable explicativa tiene con respecto al riesgo.

Por tanto, nuestro interés radica en el comportamiento de la variable Y "indicador de *cardiohealth default*" condicionada a los valores de $X = (X_1, \dots, X_j, \dots, X_p)^T$, es decir, estamos interesados en la variable condicionada $Y|X = x$, $x \in \mathbb{R}^p$. La probabilidad de la variable Y condicionada a un valor x de X, $P(Y|X = x)$ siendo $x \in \mathbb{R}^p$, llamada probabilidad a posteriori es clave para la estimación del riesgo cardiovascular ya que se define como la probabilidad a posteriori del comportamiento del paciente condicionado por la información de los factores de riesgo que lo caracterizan.

Consideramos $\Omega = \{\Pi_0, \Pi_1\}$ el espacio poblacional, denotando por Π_0 la parte de la población en la que no se ha presentado el evento cardiovascular valorado y por Π_1 el segmento de población en la que sí se ha presentado el evento cardiovascular. Haremos referencia a estas poblaciones mediante la siguiente terminología:

$$\Pi_0 = \{no\ cardiohealth\ default\} \qquad \Pi_1 = \{cardiohealth\ default\}$$

La variable aleatoria objetivo con respecto a la *probabilidad de cardiohealth default* es la variable que denominaremos *cardiohealth default* Y, variable de respuesta binaria "indicadora de la aparición del evento cardiovascular", con valores $\{0, 1\}$, que está definida sobre el espacio de probabilidad $\big(\Omega, A(\Omega), P(\cdot)\big)$ de la siguiente forma:

$$Y(w) = \begin{cases} 1, & si\ \ w \in \Pi_1 \\ 0, & si\ \ w \in \Pi_0 \end{cases} \quad , \quad para\ todo\ w \in \Omega$$

Siendo $\mathrm{A}(\Omega)$ el álgebra de sucesos de Ω y $P(\cdot)$ la probabilidad definida sobre el espacio $\big(\Omega, A(\Omega)\big)$.

La probabilidad $P(\Pi_1) = P(w \in \Pi_1 | w \in \Omega)$ que denotaremos por $P(Y = 1)$, es la probabilidad *a priori* de que un individuo de la muestra presente un evento cardiovascular, mientras que $P(\Pi_0) = P(w \in \Pi_0 | w \in \Omega)$ que denotamos por $P(Y = 0)$, es su probabilidad complementaria.

La variable Y se distribuye según una variable aleatoria de Bernouille de parámetro $p = P(Y = 1)$, $Y \sim Be(p)$, con función de probabilidad:

$$P(Y = y) = p^y(1-p)^{1-y}$$

El conjunto de características observadas sobre los individuos de la muestra constituye un vector de p variables aleatorias $X = (X_1, \dots, X_j, \dots, X_p)^T$. Si por $f_X(x)$ denotamos la función de densidad de la variable aleatoria X, y por $f_0(x)$ y $f_1(x)$ las funciones de verosimilitud de las observaciones $x \in \mathbb{R}^m$ se tiene que:

$$f_k(x) = f_{X|Y=k}(x) = \frac{P_{X,Y}\,(X = x, Y = k)}{P(Y = k)} \qquad \forall\, x \in \mathbb{R}^p, \qquad k = 0, 1 \qquad (6.1)$$

donde $P_{X,Y}\,(x, y)$ la distribución conjunta de $X \;\; e \;\; Y$.

Entonces la probabilidad del indicador de aparición del evento cardiovascular condicionada a la información proporcionada por la observación de los factores de riesgo $x \in \mathbb{R}^p$, se define como:

$$P(Y = k|X = x) = \begin{cases} \dfrac{P_{X,Y}\,(X = x, Y = k)}{f_X(x)} & \forall\, x \in \mathbb{R}^p / f_X(x) > 0 \\ 0 & e.o.c. \end{cases} \qquad (6.2)$$

Se define $P(Y = 1|X = x)$ como la *"probabilidad de cardiohealth default a posteriori"* o *"probabilidad de cardiohealth default condicionada"* por el valor conocido del vector aleatorio $x \in \mathbb{R}^p$, a la que a partir de ahora nos referiremos "*probabilidad de cardiohealth default*" o "*probabilidad de aparición del evento cardiovascular*".

Análogamente, $P(Y = 0|X = x)$ es la "*probabilidad de no cardiohealth default*" o "*probabilidad de no aparición del evento cardiovascular"*.

La ***función de probabilidad de cardiohealth default*** es por tanto, la función real con valores en $[0, 1]$ de p variables reales que asigna a cada observación $x \in \mathbb{R}^p$ la *probabilidad de cardiohealth default* o *probabilidad de aparición del evento cardiovascular*, y que denotamos por $P(Y = 1|X = x)$, para todo $x \in \mathbb{R}^p$.

A partir de (6.1) se tiene la siguiente relación fundamental que nos muestra el modelo matemático teórico que relaciona la probabilidad de *cardiohealth default* con las variables explicativas $(X_1, \dots, X_p)$:

$$P(Y=1|X=x)=\frac{P(Y=1)f_1(x)}{f_X(x)} \qquad \forall\, x\in\mathbb{R}^p$$
$$P(Y=0|X=x)=\frac{P(Y=0)f_0(x)}{f_X(x)} \qquad \forall\, x\in\mathbb{R}^p \qquad (6.3)$$

Esta relación es fundamental ya que nos muestra el modelo matemático teórico que relaciona la probabilidad de *cardiohealth default* con las variables explicativas $(X_1,\dots,X_p)$. Por esta razón a lo largo de la tesis insistiremos en el hecho de que los modelos de estimación de la probabilidad deberán preservar esta relación.

Nótese que, cuando las variables $(X_1,\dots,X_p)$ son absolutamente continuas para $x\in\mathbb{R}^p$, se tiene que $P(X=x)=0$, por lo que $P(Y=1|X=x)$ es una probabilidad condicional en sentido no elemental y debe ser tratada con cuidado.

6.2.1. La Probabilidad de *Cardiohealth Default* como Transformación Logística de la Razón de Verosimilitud

Siendo $odds(z)=\frac{z}{1-z}$, $z\in[0,1]$ y $p=P(Y=1)$, a partir de la relación (6.3) se tiene que:

$$odds\big(P(Y=1|X=x)\big)=odds(p)\times LR(x), \quad \forall x\in \text{Rango(X)}\subset\mathbb{R}^p$$

de donde,

$$log\Big(odds\big(P(Y=1|X=x)\big)\Big)=log\big(odds(p)\big)+log\big(LR(x)\big), \quad \forall x\in \text{Rango(X)}$$

Por lo que, según la notación $logit(z)=log\left(\frac{z}{1-z}\right)\ con\ z\in[0,1]$, se tiene:

$$logit\big(P(Y=1|X=x)\big)=logit(p)+log\big(LR(x)\big), \quad \forall x\in \text{Rango(X)} \qquad (6.4)$$

siendo $logit(p)$ un término constante no dependiente de x.

La parte derecha de la igualdad (6.4) es una función en x que denotaremos por $C(X)$, que contiene además de la información proporcionada por la probabilidad de *cardiohealth default* a priori, toda la información proporcionada por la razón de verosimilitud, ambas en escala logarítmica:

$$C(X) = logit(p) + log\big(LR(x)\big) \qquad (6.5)$$

La función $C(X)$ representa toda la información contenida en las variables explicativas $(X_1, \dots, X_p)$ sobre el comportamiento del paciente frente a la aparición del evento cardiovascular, por lo que en cardiología $C(X)$ se asocia con posibles patologías cardiovasculares. A la función $C(X)$ se le llamará en esta Tesis Doctoral, *función de calificación* o *función de puntuación cardiovascular*.

Una de las muchas misiones de la función $C(X)$, es pronosticar el futuro estado de Y para cada paciente a partir de la información de sus factores de riesgo cardiovascular.

La función de calificación $C(X)$ es clave para la construcción de la calificación de los pacientes, y también juega un papel importante en la construcción de la regla de decisión sobre la pertenencia o no de un paciente a la población de riesgo en un horizonte próximo.

A partir de la expresión (6.4) se puede concluir que la probabilidad de que un paciente para el que se ha observado un valor $x = (x_1, \dots, x_p)^{\mathrm{T}} \in \mathrm{Rango(X)} \subset \mathbb{R}^p$ del vector aleatorio $p - dimensional\ (X_1, \dots, X_p)^{\mathrm{T}}$, presente un accidente o un evento cardiovascular en un horizonte temporal próximo, viene dada por:

$$P(Y = 1|X = x) = \Lambda\big(C(X)\big) = \frac{1}{1 + e^{-\{logit(p) + log\left(LR(x)\right)\}}}, \quad \forall x \in Rango(X) \qquad (6.6)$$

donde $\Lambda(\cdot)$ es la función de distribución acumulada logística $L(0,1)$, $\Lambda(z) = \frac{1}{1+e^{-z}}$ para todo $z \in \mathbb{R}$, $p = P(Y = 1)$ es la probabilidad de *cardiohealth default* a priori y $LR(x)$ es la razón de verosimilitud de $\boldsymbol{X}$.

Destacar que según la relación (6.6) la probabilidad de *cardiohealth default* enlaza con las características de riesgo cardiovascular $(X_1, \dots, X_p)$ a través de la función de distribución logística, lo que nos orienta claramente sobre la parte de la estructura básica de los modelos de estimación de probabilidad de *cardiohealth default*.

La probabilidad de *cardiohealth default* puede por tanto expresarse de la forma:

$$P(Y = 1|X = x) = \Lambda\big(C(X)\big) \qquad (6.7)$$

lo que es equivalente a:

$$C(X) = logit\big(P(Y = 1|X = x)\big) \quad (6.8)$$

donde $C(X)$ viene dada por (6.5).

Las igualdades reflejadas en las expresiones (6.7) y (6.8), indistintamente constituyen un primer instrumento teórico básico para calcular la probabilidad de *cardiohealth default*.

La probabilidad con la que sea posible determinar la probabilidad de *cardiohealth default* a partir de (6.7) y (6.8), dependerá del conocimiento que se posea sobre la función $C(X) = logit(p) + log\big(LR(x)\big)$, que se situará en uno de los innumerables estadios intermedios entre las fronteras irreales del conocimiento total y el total desconocimiento.

En el supuesto de que se conozca la distribución de las probabilidades a priori de la variable estado de *cardiohealth default* Y, y las verosimilitudes del vector de variables explicativas $X = \big(X_1, \dots, X_p\big)^{\mathrm{T}}$ para las poblaciones de *cardiohealth default* y no *cardiohealth default*, la probabilidad de *cardiohealth default* quedará perfectamente determinada. Bajo la hipótesis de normalidad de las distribuciones de X condicionadas al *cardiohealth default* y no *cardiohealth default* con igual matriz de covarianzas, el $logit$ de la probabilidad de *cardiohealth default* se relaciona con el vector aleatorio X a través de la función discriminante de Fisher, frontera de clasificación del Análisis Discriminante Lineal, LDA (FISHER (1936) (54), LADD (1966) (55), LACHENBRUCH(1975) (56)).

En la práctica lo habitual es no conocer ni las probabilidades a priori ni las distribuciones de las verosimilitudes. De estas últimas con frecuencia no sólo no se conocen los parámetros sino incluso no se conoce la forma. Las probabilidades a priori pueden ser estimadas fácilmente a partir del estimador de máxima verosimilitud de la proporción de *cardiohealth default* en la muestra de pacientes, puesto que se conoce perfectamente la distribución de la variable aleatoria respuesta Y, distribución de Bernouilli de parámetro p. Si no se conoce el parámetro p, puede ser estimado paramétricamente utilizando una muestra aleatoria simple de tamaño N, $(Y_1, \dots, Y_N)$ a través de la función de verosimilitud de la muestra:

$$L(p) = p^{\sum_{i=1}^{N} y_i}(1-p)^{N-\sum_{i=1}^{N} y_i} \quad (6.9)$$

El estimador de máxima verosimilitud de p, $\hat{p}$ es la proporción de *cardiohealth default* en la muestra, que como todo estimador de máxima verosimilitud es consistente, asintóticamente eficiente y normal.

De no conocer las verosimilitudes o la razón de verosimilitud, no tendremos más remedio que estimarlas. La estimación puede llevarse a cabo mediante distintos métodos como pueden ser el método de los k vecinos más próximos o mediante funciones núcleo univariantes o multivariantes. El primero a pesar de basarse en hipótesis estructurales del modelo muy flexibles, es muy poco suave, por lo que conlleva a un sobre ajuste del modelo que resulta poco eficaz en la generalización, cualidad imprescindible para la correcta valoración del riesgo de un nuevo paciente. En relación a la estimación mediante funciones núcleo, salvo que se acepte la *hipótesis ingenua de Bayes* que asume que los elementos del vector aleatorio X condicionados a las poblaciones *cardiohealth default* y no *cardiohealth default* son independientes entre sí, se enfrenta a la maldición de la dimensionalidad. El problema es que la aceptación de la *hipótesis ingenua de Bayes* en estimación de riesgo es inaceptable. Por ello, podemos concluir que la estimación directa de las verosimilitudes sin más información que los datos de riesgo cardiovascular observados sobre los pacientes, conlleva problemas importantes que imposibilitan una solución satisfactoria.

Afortunadamente contamos con una segunda herramienta teórica para estimar la probabilidad de *cardiohealth default*, puesto que al ser la variable aleatoria respuesta Y binaria con distribución de Bernoulli de parámetro $p = P(Y = 1)$, se puede expresar la probabilidad de *cardiohealth default* a través de la esperanza condicionada:

$$P(Y = 1|X = x) = E[Y|X = x] \qquad (6.10)$$

por tanto,

$$E[Y|X = x] = \Lambda\big(C(X)\big) \qquad (6.11)$$

es decir, la función $\Lambda\big(C(X)\big)$ coincide con la función de regresión de Y sobre X, razón por la que se llama ***regresión logística*** en sentido amplio.

Dado que la esperanza condicional coincide con la regresión de Y sobre X, en el caso frecuente de no conocer el modelo teórico y todos sus parámetros, se puede trasladar toda la potencia de la regresión a la estimación de las probabilidades de *cardiohealth default*, lo que nos permitirá explicar y en su caso predecir ya sea por métodos

paramétricos, semiparamétricos o no paramétricos, el valor de Y dados los valores de X, todo ello desde la perspectiva de que nuestro objetivo será buscar los efectos de los factores simples y encontrar el mejor modelo.

El objetivo de este trabajo es **construir un modelo predictivo para estimar la *probabilidad de cardiohealth default*, $P(Y=1|X=x)$ a través de $E[Y|X=x]$**, o equivalentemente, estimar la función de regresión $r(x)=E[Y|X=x]$ a través de un modelo estadístico paramétrico o semiparamétrico, en función del grado de conocimiento real o asumido en la estructura del modelo.

En general, se buscará que el modelo exprese la relación existente entre una conveniente transformación de la probabilidad de *cardiohealth default* $g(\cdot)$, y la *función de calificación* o *función de puntuación cardiovascular* $C(x)$, función que representa toda la información contenida en las variables explicativas sobre el comportamiento del paciente frente a la aparición de un evento, es decir:

$$g\big(P(Y=1|X=x)\big)=C(x) \qquad (6.12)$$

Según sea la transformación elegida así tendremos los siguientes modelos:

Familia de Modelos	Transformación $g(\cdot)$
Probabilidad	$g\big(P(Y=1\|X=x)\big)=P(Y=1\|X=x)=C(x)$
Logísticos	$g\big(P(Y=1\|X=x)\big)=\Lambda^{-1}\big(P(Y=1\|X=x)\big)=C(x)$
Probit	$g\big(P(Y=1\|X=x)\big)=\Phi^{-1}\big(P(Y=1\|X=x)\big)=C(x)$
Vector Soporte	$g\big(P(Y=1\|X=x)\big)=Sign\left\{P(Y=1\|X=x)-\frac{1}{2}\right\}=C(x)$

siendo $\Lambda(\cdot)$ es la función de distribución acumulada logística y $\Phi(\cdot)$ la función de distribución normal estandarizada.

Una vez fijado el nexo de unión entre la *probabilidad de cardiohealth default* y la función de calificación de pacientes, estimaremos la *probabilidad de cardiohealth default* $P(Y=1|X=x)=E[Y|X=x]$, a través del correspondiente modelo de ajuste

$$g\left(\hat{P}(Y=1|X=x)\right)=\hat{C}(x)+\varepsilon \qquad (6.13)$$

Dado que la variable respuesta Y es binaria, el nexo natural que une el *verdadero modelo* de las variables explicativas con la variable repuesta es la transformación logística $g(\cdot)=logit(\cdot)$, dando origen a los modelos que nos referiremos en sentido

amplio como ***modelos logísticos***. En línea con este planteamiento analizaremos de forma destacada los modelos logísticos de expresión general,

$$logit\big(P(Y=1|X=x)\big) = C(x) \qquad (6.14)$$

y cuyo modelo de ajuste viene dado por:

$$logit\left(\hat{P}(Y=1|X=x)\right) = \hat{C}(x) + \varepsilon \qquad (6.15)$$

donde se asume que el término de error ε tiene distribución logística de media 0 y varianza $\frac{\pi^2}{3}$.

El objetivo es el de construir un modelo estadístico para alcanzar conclusiones sobre el riesgo del paciente de presentar un evento cardiovascular en un horizonte temporal próximo a través de los valores de las variables explicativas o factores de riesgo cardiovascular observados sobre el mismo. A este modelo nos referiremos como ***modelo de riesgo cardiovascular***.

Destacar que como se ha indicado en el párrafo anterior, el objetivo es el de alcanzar conclusiones sobre el riesgo del paciente, lo que conlleva un concepto más amplio que el de estimar exclusivamente la probabilidad de *cardiohealth default*. Nuestro objetivo es que el modelo sea capaz de estimar la *probabilidad de cardiohealth default* o probabilidad de presentar un evento cardiovascular y clasificar a nuevos pacientes dentro de una de las poblaciones, *cardiohealth default* o *no cardiohealth default*.

En términos de la transformación logística, a través de la probabilidad de *cardiohealth default* podría entonces obtenerse la puntuación cardiovascular asignada a cada paciente según la ecuación $C(x) = logit\big(P(Y=1|X=x)\big)$ y recíprocamente, obtenida la *función de calificación* se obtiene la probabilidad de *cardiohealth default*, $P(Y=1|X=x) = \Lambda\big(C(x)\big)$.

Además, el modelo de riesgo que se plantee debe tener en cuenta los siguientes aspectos:

1º Flexibilidad del modelo. Capacidad para describir situaciones de naturaleza diferente.

2º Dimensión del modelo. Ligada a la varianza de las estimaciones, que crece rápidamente para N fijado si p aumenta (problema de la maldición de la dimensionalidad), lo que conlleva la inestabilidad del modelo estimado.

3º Facilidad de interpretación del modelo. Este aspecto es fundamental ya que el objetivo primordial de la estimación del riesgo cardiovascular es el de su aplicación en prevención primaria. Por ello es necesario que el modelo revele la relación entre el *estado de cardiohealth default* y cada uno de los factores de riesgo cardiovascular, condicionado a la presencia del resto de factores. Además su interpretación debe ser sencilla con el objetivo de comunicar al paciente los cambios de conducta necesarios para actuar sobre los niveles de aquellos factores de riesgo que sean modificables.

4º Generalización. Es fundamental que el modelo clasifique bien no sólo a los pacientes que han servido de base para la construcción del modelo, sino que debe clasificar correctamente a nuevos pacientes, característica que se conoce como modelo generalizable o como no sobre ajustado.

El modelo ha de obtenerse de la forma más sencilla que sea posible, salvaguardando la eficacia de sus objetivos. Las descripciones deben mantenerse lo más simples posibles hasta el momento en que se demuestre que resultan inadecuadas, es el famoso principio de OCCAM, que dice que "*en igualdad de condiciones, la explicación más sencilla suele ser la más probable*". Einstein interpretó este principio diciendo, *"que el modelo sea sencillo, lo más sencillo posible, no más"*.

La estructura formal del modelo que se plantea para la *probabilidad de cardiohealth default*, $P(Y = 1|X = x),\ \forall\, x \in \mathbb{R}^p$, ha de basarse necesariamente en (6.12) y (6.13), e independientemente de la transformación $g(\cdot)$ que se adopte, la estimación de la probabilidad y de la *función de calificación* dependerá del conocimiento que se posea sobre la forma y parámetros de distribuciones de las distintas variables explicativas, ya sea de $P(X,Y)$ o bien de $P(Y = 1|X = x)$. Dependiendo de este conocimiento podremos utilizar métodos paramétricos, métodos no paramétricos o bien métodos semiparamétricos, pero en todo caso $C(X)$ es una función sobre la que realizaremos las hipótesis necesarias para obtener el "mejor modelo posible" con el fin de pronosticar el

estado de cardiohealth default o de riesgo, calificar a los pacientes y clasificar a nuevos pacientes.

El modelo quedará especificado una vez que se conozcan todos los términos de la expresión:

$$g\big(P(Y=1|X=x)\big)=C(x)$$

La especificación del modelo deberá realizarse contando con los datos disponibles y el conocimiento sobre los mismos, así como con el conocimiento del cardiólogo especialista.

6.3. REVISIÓN DE MODELOS DE ESTIMACIÓN DEL RIESGO

Una fase esencial en la construcción del modelo de riesgo cardiovascular la constituye la especificación del mismo. Basándonos en el Modelos de Probabilidad de Default Generalizado, GDPM, se buscará el modelo que mejor exprese la relación existente entre una conveniente transformación de la *probabilidad de cardiohealth default*, $g(\cdot)$, y la *función de calificación* o *función de puntuación cardiovascular* $C(x)$, es decir especificar la estructura formal (6.12) a través de una relación de la forma:

$$g\big(P(Y=1|X=x)\big)=C(x)=\sum_{r=0}^{q}\beta_r h_r(X)=\boldsymbol{\beta^T H(X)} \qquad (6.16)$$

siendo:

$h_r(X)$ la $r-ésima$ transformación de X, llamada $r-ésima$ función de base de X, $r=0,\dots,q$

$\sum_{r=0}^{q}\beta_r h_r(X)=\boldsymbol{\beta^T H(X)}$ una expansión lineal por funciones de base del vector de variables explicativas $\boldsymbol{X}=\big(X_1,\dots,X_p\big)^T$ y $\boldsymbol{\beta}=\big(\beta_1,\dots,\beta_q\big)^T$ vector q – dimensional de coeficientes desconocidos.

Este planteamiento añade la no linealidad a los modelos utilizando expansiones lineales de las variables explicativas que tienen como idea básica, sugerida por HASTIE y TISBSHIRANI (1996) (57), reemplazar el vector de variables explicativas $\boldsymbol{X}$ por funciones de base, las cuales son transformaciones de $\boldsymbol{X}$, permitiendo así construir un modelo más flexible y así emplear modelos lineales en este nuevo espacio. Estos

espacios *agrandados* de las expansiones por funciones de base de las variables explicativas son generalmente espacios de Hilbert. El mejor acontecimiento en la especificación del modelo es encontrar las adecuadas funciones de base $\left(h_0(X), \dots, h_q(X)\right)^T$.

Por tanto, la estructura formal del modelo (6.16) quedará perfectamente determinada una vez fijada la función de enlace $g(\cdot)$ y la expansión lineal de funciones de base, $\sum_{r=0}^{q} \beta_r h_r(X)$.

En general, esta selección depende principalmente del grado de conocimiento que se posea sobre la relación de dependencia del *estado de cardiohealth default* y las variables explicativas o factores de riesgo cardiovascular. Este conocimiento puede situarse entre el total desconocimiento, en el que se supone no conocida la distribución poblacional ni ninguna estructura que refleje la relación de dependencia entre la variable *estado de cardiohealth default* y los factores de riesgo; y el conocimiento total, que sería el extremo opuesto en el que se supone que conocemos la distribución conjunta de las variables explicativas y la variable respuesta, $P(X, Y)$ así como la *probabilidad de cardiohealth default a priori*, $P(Y = 1)$.

El modelo propuesto no parte de ninguno de estos dos extremos de conocimiento, desconocimiento total y conocimiento total, sino que de una situación intermedia y adaptada a la realidad del contexto en el que nos encontramos, y es el de un modelo basado en el conocimiento parcial de $C(x)$.

Así podremos especificar en el modelo de manera fundamentada, algunas o todas las variables con estructura lineal o con estructura no lineal, y dentro de la no linealidad seleccionar algunas de las innumerables formas en que ésta puede manifestarse, todo ello estructurado sobre el planteamiento común de las expansiones lineales por funciones de base de las variables explicativas del *estado de cardiohealth default*.

En lo que respecta a la especificación de la función de enlace en la relación (6.12), en un principio se podrían proponer modelos donde $g(\cdot)$ fuera la función logística, la función probit o la función vector soporte. Según la selección se tendrían los modelos logísticos, probit y vector soporte respectivamente. Estos últimos se ha decidido descartarlos ya que no proporcionan directamente estimadores de la *probabilidad de cardiohealth default* y estarían orientados solamente a la clasificación(53). Además,

dado que la variable respuesta Y es binaria, el nexo natural que une el *verdadero modelo* de las variables explicativas con la variable repuesta es la transformación logística $g(\cdot) = logit(\cdot)$. Por ello, la propuesta se basará en la función logística como función de enlace, dando origen a los modelos que según hemos mencionado nos referiremos en sentido amplio como ***modelos logísticos***

A continuación, se revisan los distintos modelos propuestos por F. Mallo Fernández en su Tesis Doctoral titulada *"Modelos Multivariantes Internos de Medición de Riesgos de Crédito, acordes a Basilea II"* (2011) (53) destacando sus debilidades y fortalezas para así poder concluir con el mejor modelo de estimación del riesgo en nuestro campo.

Suponemos que el conocimiento que se posee sobre la relación de dependencia del estado de *cardiohealth default* con las variables explicativas nos lleva a modelos basados en el conocimiento parcial de la *función de calificación*, suposición que consideramos la más razonable de aquellas que se sitúan entre el desconocimiento total y el conocimiento total.

Los modelos disponibles para representar la relación de dependencia entre el logit de la probabilidad de *cardiohealth default a posteriori* y las variables explicativas del riesgo cardiovascular, y las técnicas para estimarlos, se pueden clasificar en función de que el modelo contemple la linealidad o no linealidad de las variables que lo conforman.

6.3.1. Regresión Logística Lineal, LLR o LOGIT

Cuando hay razones suficientes para suponer que todas las variables explicativas son lineales, o que al menos este supuesto puede recoger adecuadamente la relación de dependencia entre la variable estado de *cardiohealth default* y las variables explicativas, se establece una hipótesis muy simple, la *función de calificación* $C(X)$ es lineal en $\boldsymbol{X} = (X_1, \dots, X_p)^{\mathrm{T}}$.

Bajo la hipótesis de linealidad y considerando la función de enlace logística, el modelo presenta la siguiente estructura formal:

$$logit\big(P(Y = 1|X = x)\big) = \beta_0 + \sum_{i=1}^{p} \beta_r X_r = \beta_0 + \boldsymbol{\beta^T X} \qquad (6.17)$$

donde $\boldsymbol{X} = (X_1, \dots, X_p)^T$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$.

El modelo (6.17) es conocido como *LOGIT* y su ajuste se realiza mediante la ***Regresión Logística Lineal, LLR***.

La estimación de los parámetros del modelo (6.17) se lleva a cabo mediante la minimización del riesgo empírico, lo que puede conllevar a problemas tales como infinitas soluciones y sobre e infra ajuste. Esta problemática podría resolverse activando un término de regularización y resolviendo el problema de minimización correspondiente. En este último caso nos encontraríamos ante la ***Regresión Logística Lineal L_2 – Penalizada***.

Aunque la linealidad es una característica muy deseable, no siempre es alcanzable. En los casos en los que no se alcanza, situación que ocurre con frecuencia en la práctica, son necesarios métodos alternativos a los modelos lineales, pero para ello debemos de enfrentarnos al dilema de Occam con decisión, lo que significa obtener el modelo de la forma más sencilla que sea posible salvaguardando la eficacia de los objetivos perseguidos.

A continuación se presenta la revisión de modelos de estimación bajo la suposición de no linealidad de algunas variables explicativas.

6.3.2. Modelos Aditivos Generalizados, GAM

Una familia de modelos para caracterizar la no linealidad de las variables explicativas a la vez que eliminar la maldición de la dimensionalidad es la de los Modelos Aditivos Generalizados, *GAM*. Estos modelos son extensiones de los **Modelos Lineales Generalizados, *GLM***, que combinan la precisión estadística típica de una variable explicativa unidimensional con la flexibilidad de los modelos semiparamétricos de variables explicativas multidimensionales, por lo que en estos modelos no está presente la maldición de la dimensionalidad.

La *función de calificación* adopta la forma $C(X) = \beta_0 + \sum_{r=1}^{q} \beta_r h_r(X)$, la función de enlace viene dada por $g(\cdot)$ y las funciones de base por:

$$h_r(X) = C_r(X_r), \text{ función de suavizado de dimensión infinita}$$
$$E[C_r(X_r)] = 0, \;\; r = 1, \dots, p \qquad \text{(6.18)}$$

Por tanto, el modelo presenta la siguiente estructura formal:

$$g\big(P(Y=1|X=x)\big) = \beta_0 + \sum_{i=1}^{p} C_r(X_r) \ , \quad \beta_0 \in \mathbb{R} \qquad (6.19)$$

El planteamiento de Modelos Aditivos Generalizados presenta las siguientes ventajas:

- Cada uno de los términos aditivos se estima usando un suavizador univariante, por lo que se esquiva la maldición de la dimensionalidad.
- Las estimaciones de los términos individuales explican cómo cambia la variable respuesta con las correspondientes variables explicativas observadas, lo que confiere a los modelos aditivos la habilidad de descubrir los patrones no lineales sin sacrificar la interpretabilidad.

La facilidad de interpretación hace que los modelos *GAM* sean particularmente atractivos en los sistemas de calificación del riesgo cardiovascular, porque no olvidemos que uno de los objetivos de la estimación es el de transmitir al paciente su riesgo cardiovascular y plantear una modificación de conducta para aquellos factores de riesgo que sean modificables.

Dentro de la familia de los Modelos Aditivos Generalizados destacamos particularmente dos:

- Modelo de ***R**egresión **L**ogística **A**ditiva **(ALR)***, en el que la función de enlace es la función logística $\big(g(\cdot) = logit(\cdot)\big)$. Las funciones $C_r(X_r)$ pueden estimarse de una manera flexible a través de cualquier método de suavizado no paramétrico.
- ***R**egresión **L**ogística **A**ditiva **R**egularizada **(RALR):*** La función de enlace viene dada por la función logística, $g(\cdot) = logit(\cdot)$, y se especifica la función de suavizado de dimensión infinita $C_r(x_r)$ como una expansión de una colección arbitrariamente larga de funciones de base, controlando la complejidad a través del regularizador $J(C(X))$.

Tanto *ALR* como *RALR* proporcionan directamente la *probabilidad de cardiohealth default*, y por tanto, la función de calificación correspondiente. Además aceptan de forma flexible la no linealidad y permiten conocer los efectos marginales de las distintas variables de entrada asociadas con el riesgo.

A pesar de las ventajas mencionadas, los principales inconvenientes que presentan son:

- La dificultad de interpretación de las funciones paramétricas de dimensión infinita, que por ello lo convierte en poco adecuados para la estimación del riesgo cardiovascular en el que la facilidad de interpretación del modelo es un requisito.
- Eluden la maldición de la dimensionalidad suponiendo la aditividad de los efectos de riesgo en su relación con el *estado de cardiohealth default*, y esta hipótesis de aditividad es comprometida ya que es poco realista en la estimación del riesgo cardiovascular.

6.3.3. Árboles de decisión, TREE

Una técnica para especificar y caracterizar la no linealidad de las variables explicativas en los modelos estadísticos, con un planteamiento mucho menos comprometido que la exigencia de aditividad de los efectos de las variables explicativas y mucho más simple en su desarrollo teórico, consiste en dividir el espacio de características en un conjunto de híper – rectángulos, que constituyen una partición recursiva y entonces ajustar un simple modelo (igual a una constante) en cada uno, obteniendo de este modo una regla de predicción de la *probabilidad de cardiohealth default*.

Los árboles de decisión son conceptualmente simples y muy atractivos, ya que están dotados de una gran facilidad de interpretación. Además tienen otra gran ventaja y es que detectan de forma automática estructuras complejas entre variables.

A pesar de las ventajas descritas, este tipo de modelos no son adecuados en nuestro caso ya que además de su alta varianza, la probabilidad estimada es constante a trozos, característica que no se ajusta dada la forma usual de la función subyacente real.

6.3.4. Splines de Regresión Adaptativos Multivariantes, MARS

Las técnicas de construcción y estimación de los modelos de regresión *MARS* son técnicas semiparamétricas adaptativas, generalización de los modelos lineales paso a paso que modelan de forma automática relaciones no lineales e interacciones entre la variable respuesta y las variables explicativas. El modelo resultante es un modelo continuo muy flexible obtenido por ajuste de regresiones lineales a trozos, con derivadas continuas, muy eficaz para encontrar a través de transformaciones óptimas de las variables y sus interacciones, la compleja estructura que frecuentemente se esconde en los datos.

A pesar de ser un modelo bastante eficaz, se descarta puesto que no tiene en cuenta que la variable respuesta *estado de cardiohealth default* es binaria, por lo que no restringe los valores ajustados entre cero y uno, por lo que tampoco se ajusta correctamente dada la forma usual de la función subyacente real.

6.3.5. Modelo Perceptron de Capa Simple Oculta, SLPM

En la misma línea que *PPLR* y con una estructura funcional muy similar se mueve el *Modelo Perceptron de Capa Simple* de transmisión de información hacia adelante y una sola capa oculta, *SLPM*, una técnica de la familia de Redes Neuronales Artificiales, ANNs, que conecta a esta familia con los Modelos Lineales de Funciones de Base.

Las redes neuronales no gozan en general de buena fama en el campo de la predicción, ya que ha existido un gran número de de redes neuronales de tipo publicitario que han conseguido que estos modelos sean vistos como cajas negra mágicas y misteriosas y en la mayor parte de los casos así es.

En los sistemas de calificación de los acreditados se han usado modelos de redes neuronales pero más orientados a la decisión sobre la concesión de un crédito que a estimar un modelo de probabilidad concebido desde la óptica de los acuerdos de Basilea II, en ese sentido, aparte de que no proporcionan en muchos casos la *probabilidad de default*, constituyen siempre una "caja negra" incapaz de explicar la relación entre el *default* y las variables explicativas del riesgo. En nuestro caso, el poder explicar la relación entre el estado de *cardiohealth default* y los factores de riesgo es imprescindible, por ello este tipo de modelos se ha descartado.

6.3.6. Modelos Regularizados por Núcleos, KRM

A través de las expansiones lineales por funciones de base de las variables explicativas se trasladan los datos a un espacio agrandado, generalmente de Hilbert, $\mathcal{H}$. Con la frontera lineal en el espacio agrandado se consigue una mejor separación de los datos de entrenamiento, separación que se traslada a la frontera no lineal en el espacio original

El peligro está en que con las suficientes expansiones de funciones de base los datos pueden hacerse separables de forma artificial, resultando sobre ajustados. Una alternativa consiste en extender la idea anterior a espacios de alta dimensión y controlar la complejidad del modelo ajustando la función con el criterio de minimización de la pérdida empírica asociada a una cierta función de pérdida regularizada. Los métodos

que estiman el modelo de acuerdo con esta filosofía se llaman Métodos Regularizados por Núcleos, *KRM*.

Los modelos *KRM* se basan en la idea de agrandar el espacio usando para la *función de calificación* expansiones por funciones de base de las variables originales. Además aprovechan el hecho de que la formulación del riesgo empírico admite una gran flexibilidad para la *función de calificación* $C(X)$, con generalizaciones no lineales como $C(X) \in \mathcal{H}$, siendo $C(X)$ una función arbitraria y $\mathcal{H}$ un espacio de Hilbert.

Si la función de enlace que se considera es la logística, $g(\cdot) = logit(\cdot)$, se tiene el **M**odelo de ***R***egresión ***L***ogística ***R***egularizada por ***N***úcleos, ***KLRM***. En este modelo se estima directamente la *probabilidad de cardiohealth default*, pero no se resuelve el problema de la interpretabilidad, ya que describe la relación de dependencia de la variable estado de *cardiohealth default* con los pacientes y no con los factores de riesgo cardiovascular.

6.3.7. Modelos Parcialmente Lineales, LPM

Este tipo de modelos se encuentran en una situación intermedia entre los modelos totalmente lineales y los descritos hasta ahora en el apartado 6.3. Se configuran considerando como hipótesis de partida que se tiene información fundada sobre el hecho de que una o varias de las variables de interés tienen influencia lineal sobre el comportamiento frente al *cardiohealth default* y el resto influencia no lineal.

Suponiendo que tenemos p_1 variables lineales, $(X_1, \dots, X_{p_1})$ y $p_2 = p - p_1$ variables de las que se desconoce el tipo de influencia que ejercen. Dado que $X^T = (X_1, \dots, X_{p_1}, X_{p_1+1}, \dots, X_{p_2})$, adoptando la notación:

$$U^T = (U_1, \dots, U_{p_1}), \text{ siendo } U_j = X_j \text{ para } j = 1, \dots, p_1$$

$$V^T = (V_1, \dots, V_{p_2}), \text{ siendo } V_j = X_{p_1+j} \text{ para } j = 1, \dots, p_2$$

Se tiene que $X^T = (U^T, V^T)$, notación mediante la cual expresaremos la estructura formal del modelo como la suma de una componente lineal, combinación de las variables lineales $U^T = (U_1, \dots, U_{p_1})$, y una componente no lineal, que se diseña como una expansión no paramétrica $h(\boldsymbol{V})$, donde $h(\cdot)$ es una función no paramétrica infinito dimensional.

El modelo presenta la siguiente estructura formal:

$$g\big(P(Y=1|X=x)\big)=\beta_0+\sum_{r=1}^{p_1}\beta_r U_r+h\big(V_1,\dots,V_{p_2}\big)=\boldsymbol{\beta}^T\boldsymbol{U}+h(\boldsymbol{V}) \qquad (6.20)$$

donde $h(\boldsymbol{V})$ es una función no paramétrica infinito dimensional y $\beta_{p_1+1}=\cdots=\beta_p=1$.

Si se supone la función de enlace logística, $g(\cdot)=logit(\cdot)$ y $V^T=(V_1,\dots,V_{p_2})$ absolutamente continuas, estamos ante el ***M**odelo **L**ogístico **P**arcialmente **L**ineal,* ***LPLM***.

$$logit\big(P(Y=1|X=x)\big)=\beta_0+\sum_{r=1}^{p_1}\beta_r U_r+h\big(V_1,\dots,V_{p_2}\big)=\boldsymbol{\beta}^T\boldsymbol{U}+h(\boldsymbol{V}) \qquad (6.21)$$

En este tipo de modelos se estiman simultáneamente las puntuaciones de riesgo cardiovascular y las *probabilidades de cardiohealth default*. Además permiten captar la linealidad paramétricamente, consideran la componente no lineal y no son excesivamente complejos. El principal inconveniente es que la parte no lineal se estima a través de una función no paramétrica infinito dimensional, que generalmente no permite explicar la aportación de cada variable explicativa del riesgo al *estado de cardiohealth default*. Además, como ocurre con casi todos los métodos no paramétricos, estas técnicas están aquejadas por la maldición de la dimensionalidad. Por esta razón este tipo de modelos no se consideran satisfactorios para nuestro objetivo de estimar el riesgo cardiovascular, clasificar atendiendo al nivel de riesgo y conocer la influencia que cada variable explicativa tiene con respecto al riesgo.

Con el fin de resolver el problema anterior, surgió una familia de modelos logísticos parcialmente lineales, los ***M**odelos **L**ogísticos **A**ditivos **P**arcialmente **L**ineales,* ***LPALM***. Este tipo de modelos pueden considerarse desde dos ópticas, como Modelos Logísticos Aditivos con una Componente Lineal o bien como una extensión de los Modelos Logísticos Lineales con una Componente No Lineal Aditiva. La estructura formal de un *LPALM* viene dada por la expresión (6.20) considerando la función de enlace logística y donde la función no paramétrica $h(\boldsymbol{V})$ de dimensión infinita consiste en una expansión aditiva de funciones de base $h_r(V_r)$ de las variables con influencia no lineal, es decir, el modelo presentaría la siguiente estructura formal:

$$logit\big(P(Y=1|X=x)\big) = \beta_0 + \sum_{r=1}^{p_1} \beta_r U_r + \sum_{r=p_1+1}^{p} h_r(V_r) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{H}(\boldsymbol{V}) \qquad (6.22)$$

En este modelo cada término aditivo se estima mediante un suavizador univariante, por lo que se esquiva la maldición de la dimensionalidad. Además las estimaciones de los términos individuales explican cómo cambia la variable respuesta con las correspondientes variables explicativas del riesgo observadas. Por ello, estos modelos son especialmente atractivos en los sistemas de calificación del riesgo, ya que la facilidad o no de interpretación dependerá de las funciones de base $h_r(\boldsymbol{V_r})$ utilizadas.

El principal problema que presenta y motivo por el cual este modelo se aleja de nuestro objetivo, es la hipótesis de la aditividad, que es muy comprometida.

6.4. SELECCIÓN DEL MODELO DE ESTIMACIÓN DEL RIESGO CARDIOVASCULAR

En el apartado *6.3 REVISIÓN DE MODELOS DE ESTIMACIÓN DEL RIESGO* y con el fin de especificar el modelo que mejor exprese la relación existente entre una conveniente transformación de la *probabilidad de cardiohealth default* y la *función de calificación*, es decir especificar la estructura formal (6.12) se han planteado distintos modelos basándonos en el Modelo de Probabilidad de Default Generalizado. Los modelos propuestos no son óptimos puesto que presentan una serie de debilidades que hace que se descarte su elección. En general, las limitaciones encontradas se pueden resumir en:

- La suposición de aditividad de los efectos de riesgo en su relación con el *estado de cardiohealth default*.
- La estructura del modelo no tiene en cuenta la naturaleza binaria de la variable respuesta *estado de cardiohealth default*.
- Las variables con un comportamiento no lineal se especifican de manera homogénea *(PPLR)*. El especificar la no linealidad para todas las variables que lo requieren del mismo modo no permite especificar las distintas manifestaciones de la no linealidad.
- Maldición de la dimensionalidad.

- La dificultad de interpretación del modelo estimado.
- La estimación de la probabilidad no se ajusta al concepto formal de la misma, bien por ser una función constante a trozos *(TREE)* o bien porque no restringe sus valores entre cero y uno *(MARS)*.

El *Modelo Logístico Parcialmente Lineal, LPLM,* que capta la linealidad paramétricamente, estima simultáneamente las puntuaciones de riesgo cardiovascular y las *probabilidades de cardiohealth default*. El principal inconveniente es que la parte no lineal se estima a través de una función no paramétrica infinito dimensional, que generalmente no permite explicar la aportación de cada variable explicativa del riesgo al *estado de cardiohealth default*, y además como ocurre con la mayor parte de los métodos no paramétricos, esta técnica está aquejada de la maldición de la dimensionalidad.

El *Modelos Logísticos Aditivos Parcialmente Lineales, LPALM*, cuya estructura formal viene dada por (6.22), donde la función no paramétrica $\boldsymbol{H}(\boldsymbol{V})$ de dimensión infinita se sustituye por una expansión aditiva de funciones de base $h_r(\boldsymbol{V_r})$. La componente no lineal se constituye así en una estructura aditiva de los efectos no lineales. Por tanto, *LPALM* trata de resolver los problemas presentados por *LPLM* a través de la estimación mediante suavizadores univariantes, por lo que se esquiva la maldición de la dimensionalidad. Además las estimaciones de los términos individuales explican cómo cambia la variable respuesta con las correspondientes variables explicativas del riesgo observadas. El principal inconveniente que presenta es que parte del supuesto de aditividad.

Ante esta situación, es necesario contar con técnicas alternativas con al menos una serie de cualidades que sean una combinación adecuada de propiedades desde el punto de vista del riesgo cardiovascular y desde el rigor estadístico, y que son las siguientes:

i. Desde el punto de vista del riesgo cardiovascular, el modelo debe responder al triple objetivo de estimar la *probabilidad de cardiohealth default* o probabilidad de presentar un evento cardiovascular, estimar la *función de calificación* de los pacientes y clasificar a nuevos pacientes dentro de una de las poblaciones, *cardiohealth default* o *no cardiohealth default*.

La combinación de los tres requerimientos del párrafo anterior determina todas las propiedades que ha de tener el modelo más idóneo para cada situación concreta.

- Equilibrio entre la capacidad del modelo para describir situaciones de naturaleza diferente.
- La complejidad del modelo y la dimensión del mismo.
- El modelo debe extender bien tal relación de dependencia (generalización) así como clasificar correctamente a nuevos pacientes.

ii. El rigor estadístico nos induce a considerar prioritariamente la función de distribución logística como enlace entre la *probabilidad de cardiohealth default* o probabilidad de presentar un evento cardiovascular y la *función de calificación*.

Las condiciones anteriores nos sitúan en una clase de modelos donde se combinen ideas de los Modelos Logísticos Aditivos Parcialmente Lineales, *LPALM*, que conservando la facilidad de interpretación introducen la flexibilidad necesaria para contemplar la no linealidad, y de modelos logísticos expansiones lineales de funciones de base de la matriz de datos original, cuya característica más destacable es que es posible asignar a cada variable no lineal combinaciones lineales de funciones de base para especificar la no linealidad. Una vez que las funciones de base (nuevas variables) han sido determinadas, los modelos son lineales en estas nuevas variables, resultando una familia de modelos logísticos lineales dentro de la cual situamos nuestra propuesta de **M**odelos **L**ogísticos **L**ineales por expansiones lineales **H**íbridas de funciones de base, *HLLM*.

La estructura funcional del modelo logístico de probabilidad por expansiones lineales de funciones de base será:

$$logit\big(P(Y=1|X=x)\big) = \beta_0 + \sum_{r=1}^{q} \beta_r h_r(X) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{H}(X) \qquad (6.23)$$

Si se consideran las variables explicativas de riesgo cardiovascular, $X^T = (U^T, V^T)$, con $U^T = (U_1, \dots, U_{p_1})$ vector de variables lineales y $V^T = (V_1, \dots, V_{p_2})$ vector de variables no lineales, donde la expansión lineal de funciones de base de las variables explicativas del riesgo es suma de una componente lineal, $\beta_0 + \sum_{r=1}^{p_1} \beta_r U_r$, y una estructura aditiva

de p_2 funciones de las variables no lineales, $Z_r(V_r)$, para cada $r = 1, \dots, p_2$, $\sum_{r=1}^{p_2} Z_r(V_r)$.

El número de funciones de base que integran la combinación lineal $Z_r(V_r)$, dependerá del tipo de la relación de dependencia no lineal entre el estado de *cardiohealth default* y la variable V_r así como del método utilizado para captarla, por lo que para cada $r = 1, \dots, p_2$, $Z_r(V_r)$ se puede expresar en la forma siguiente:

$$Z_r(V_r) = \sum_{k=1}^{K_r} \theta_{r,k}\, h_{r,k}(V_r), \quad r = 1, \dots, p_2 \qquad (6.24)$$

siendo K_r el número de funciones de base de la combinación lineal $Z_r(V_r)$ y $h_{r,k}(\cdot)$ su $k-ésima$ función de base.

La estructura formal del modelo (6.23) según la propuesta se denomina ***Modelo Logístico Lineal Híbrido, HLLM***, y se expresa de la forma:

$$logit\big(P(Y=1|X=x)\big) = \beta_0 + \sum_{r=1}^{p_1} \beta_r U_r + \sum_{r=1}^{p_2} \left(\sum_{k=1}^{K_r} \theta_{r,k}\, h_{r,k}(V_r) \right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{\theta}^T \boldsymbol{H}(\boldsymbol{V}) \qquad (6.25)$$

donde $p_1 + \sum_{r=1}^{p_2} K_r = q$ y $p_1 + p_2 = p$, $\boldsymbol{X}^T = (\boldsymbol{U}^T, \boldsymbol{V}^T)$, con $\boldsymbol{U}^T = (U_1, \dots, U_{p_1})$ vector de variables lineales y $\boldsymbol{V}^T = (V_1, \dots, V_{p_2})$ vector de variables no lineales, $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_{p_1})$, $\boldsymbol{H}(\boldsymbol{V}) = \left(\boldsymbol{H}_1(V_1), \dots, \boldsymbol{H}_{p_2}\left(V_{p_2}\right)\right)$, $\boldsymbol{H}_r(V_r) = \left(h_{r,1}(V_r), \dots, H_{r,K_r}(V_r)\right)$ con $r = 1, \dots, p_2$, $\boldsymbol{\theta}^T = (\theta_1, \dots, \theta_{p_2})$ y $\boldsymbol{\theta}_r^T = (\theta_{r,1}, \dots, \theta_{r,K_r})$.

El calificativo *híbrido* se debe a que cada variable no lineal V_r, $r = 1, \dots, p_2$, puede ser expresada como una combinación lineal de funciones de base específicas diferentes, como por ejemplo transformaciones polinómicas, logarítmicas, pesos de la evidencia obtenidas por un proceso de tramado óptimo de la variable V_r, funciones constantes a trozos resultados de particiones recursivas de árboles de decisión, funciones lineales a trozos, splines de regresión, de suavizado o de penalización, funciones sierra obtenidas por el método *PPR* sobre la variable V_r, funciones bisagra obtenidas por el procedimiento *MARS* sobre V_r, funciones base radial Gaussiana, etc.

Una vez que se han determinado para cada variable no lineal V_r, las funciones de base que forman la combinación lineal (6.24), el complejo modelo no lineal (6.20) se

derrumba dando paso al nuevo modelo lineal en el espacio expansionado de las funciones de base.

7. CARDIOVASCULAR RISK SCORECARD

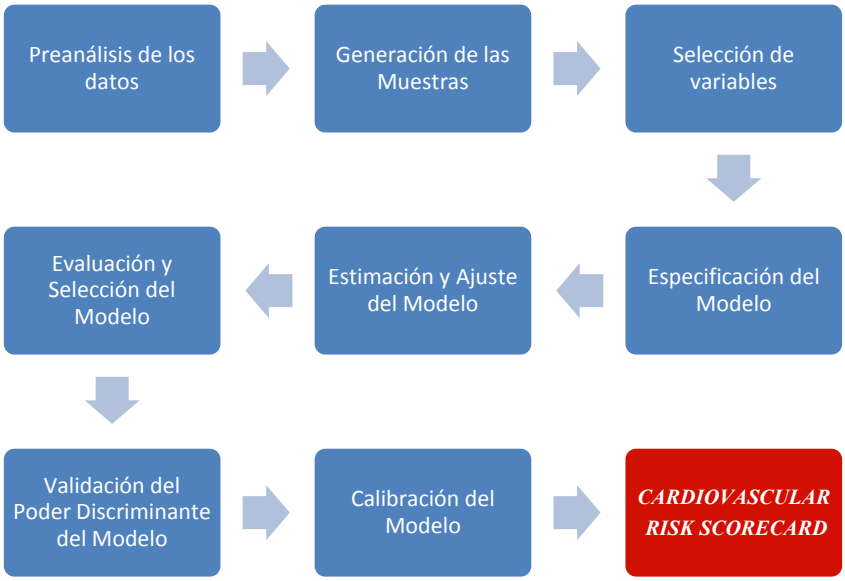
7.1. INTRODUCCIÓN

Este capítulo está dedicado a la construcción de un sistema de cuantificación del riesgo cardiovascular de pacientes a través de los Modelos Logísticos Lineales Híbridos por Expansiones Lineales de Funciones de Base, *HLLM* propuestos en el capítulo 6 y que llamaremos ***Spanish Cardiovascular Risk Scorecard***.

La propuesta que se plantea en esta Tesis marca un punto de inflexión con respecto a las tablas de riesgo cardiovascular actuales. Esto se debe por un lado, al modelo estadístico seleccionado para realizar las estimaciones de la probabilidad de *cardiohealth default*, siendo éste un modelo logístico por expansiones de base. Además, la propuesta resuelve limitaciones que presentan las tablas de predicción del riesgo cardiovascular de las que se dispone en la actualidad. Las tablas de cuantificación del riesgo cardiovascular actuales, estiman la probabilidad de presentar el evento cardiovascular valorado con un amplio horizonte temporal, que habitualmente es de diez años. Este planteamiento a largo plazo conlleva una falta de precisión en las estimaciones. Esta falta de precisión se debe a que la predicción sobre la aparición o no del evento cardiovascular valorado se realiza a partir de los valores actuales de las variables explicativas del riesgo, sin ningún tipo de corrección por la modificación que dichos valores puedan sufrir a lo largo de los diez años del horizonte considerado. En cambio, nuestra propuesta sugiere predicciones con un horizonte temporal de un año, momento en el cual el modelo se calibrará para así tener asegurada la precisión de las estimaciones que con el paso del tiempo se realicen.

Spanish Cardiovascular Risk Scorecard se diseña como un método teórico general para la estimación del riesgo cardiovascular. Se detallan las fases de las que consta la construcción del modelo de riesgo, proponiendo en cada una de ellas la metodología a seguir con el objetivo de obtener un modelo interpretable, con el triple objetivo de predecir, calificar y clasificar a nuevos pacientes susceptibles de recibir una valoración de su riesgo cardiovascular. A partir de este *modelo óptimo* que será calibrado anualmente, será posible calcular la probabilidad de *cardiohealth default a posteriori*, clave en el modelo de riesgo a través de la que se evaluará la salud cardiovascular del paciente. Igualmente, el modelo facilita la estimación de la función de calificación de pacientes y el clasificador de nuevos pacientes susceptibles de recibir una valoración de su riesgo cardiovascular.

Por tanto, nuestro objetivo es el de generalizar la construcción del sistema de calificación del riesgo cardiovascular de un modo global a través de lo que denominaremos ***Algoritmo General de Construcción*** **del** ***Spanish Cardiovascular Risk Scorecard*** y que consta de las siguientes fases:



La forma en que se acometen la mayor parte de las fases de construcción de un modelo de riesgo depende fundamentalmente de su estructura formal y de la metodología que se adopte para la estimación, evaluación, validación y calibración del mismo. Existen varios métodos para este cometido, y tenemos interés en aquellos modelos estadísticos o algoritmos de aprendizaje que además de establecer y cuantificar la relación entre las características de riesgo cardiovascular observadas sobre los pacientes y el comportamiento de éstos frente a la variable estado de *cardiohealth default*, permitan la evaluación y validación necesarias para asegurar la generalización del modelo, así como su calibración para detectar a posteriori la eficacia del modelo en sus pronósticos. Además, el modelo deberá explicar suficientemente la relación de dependencia entre el estado de *cardiohealth default* y los factores de riesgo cardiovascular que hayan sido seleccionados para formar parte del mismo.

La metodología que se va a presentar parte de la suposición de riqueza de datos de los que se dispone para llevar a cabo la estimación.

7.2. PREANÁLISIS DE LOS DATOS

En esta fase se profundiza en las características más importantes de las variables, desde su especificación, tipo de variable y rol específico hasta su significado e importancia para medir o explicar el riesgo cardiovascular.

Se realizará un análisis estadístico individual para cada una de las variables con el fin de resolver aspectos tales como el problema en su caso de de datos faltantes y la eliminación de los valores extremos que pudieran distorsionar la construcción del modelo. Para resolver estas situaciones se plantea lo siguiente:

- Problema de datos faltantes:

Un modelo paramétrico ajusta los parámetros correspondientes a partir de los casos para los que la información está completa. Lo que significa que ajusta los parámetros exclusivamente con los casos donde ninguna de las variables analizadas contiene valores ausentes. Por ello, se deben tomar medidas correctoras con el fin de maximizar el número de registros completos.

En el caso de pocas faltas, se propone una solución a través de los métodos clásicos. Para variables continuas, a los datos faltantes equivalentes a cero se les imputará el valor medio de los valores informados o la mediana, que está menos influenciada por valores extremos. Y si la variable es categórica, se les imputará la moda de los valores informados ya que no tiene sentido analizar su equivalencia con el valor cero. En caso de que no sean pocas las faltas, la solución pasa por asignar el peso de la evidencia, *WOE*, a las categorías resultantes de un proceso de discretización optima, inicialmente automático y posteriormente adaptado con criterios de riesgos cardiovascular.

- Eliminación de los valores extremos:

La presencia de valores extremos puede ser debida a errores producidos en el momento de la extracción de los datos o a características excepcionales de los registros seleccionados que presenten valores muy poco frecuentes. Estos casos deben eliminarse de la muestra de análisis ya que pueden distorsionar el modelo final. Si se incluyera un número elevado de valores extremos el modelo resultante podría estar sobre ajustado y que por tanto, funcionara bien sobre la muestra particular pero que no generalizara bien.

La solución que se plantea es la de eliminar los casos extremos bajo el principio general de que no se deben quitar demasiados casos malos ni demasiados casos buenos de la muestra.

A continuación se estudiará la relación de todas y cada una de las variables con el resto así como su significado y su significado en términos de poder de explicativo. Esto conlleva analizar cuestiones tales como la correlación, la multicolinealidad o la asociación parcial con el estado de *cardiohealth default*.

a) La detección de la correlación se llevará a cabo utilizando la matriz de correlaciones del análisis de la correlación bivariante. En el caso de ser necesario el agrupamiento de las variables explicativas en clases de variables caracterizadas por su interrelación utilizaremos una variante del Análisis de Componentes Oblicuas.

b) El análisis y la eliminación de la multicolinealidad entre las variables es clave con el fin de evitar entre otros problemas, que pequeñas variaciones de los datos provoquen cambios significativos en los coeficientes del modelo. Para ello se plantea utilizar el *Factor de Inflación de la Varianza* (BELSEY et al. (1980) (58), y KLEINBAUM et al. (1988) (59)), y los *Índices de Condición* y *Proporción de Varianza*. Los dos últimos indicadores se obtienen a partir de los valores propios de la matriz de correlaciones entre las variables independientes.

c) Un aspecto clave en esta fase es el de detectar las variables con mayor potencial de información tanto en cantidad como en calidad en relación con el estado de *cardiohealth default* de los pacientes. El poder explicativo de la variable se medirá a través de la cantidad de información proporcionada por ésta, para lo que se plantea utilizar los *estadísticos de Gini* y *Valor de la Información*, (KULBACK, 1959 (60)). La asociación entre la variable explicativa y el estado de *default* se mide utilizando test de asociación con la variable respuesta, en función de que las variables sean de intervalo o nominales, descartando aquellas variables para las que la cantidad de información no sea suficiente y/o los test de asociación no indiquen asociación.

Este proceso implica una reducción de variables, descartando aquellas que mostrando colinealidad con otras poseen escaso poder explicativo sobre el estado de default o que presenten escasa asociación con el mismo. Mediante la reducción de variables en los

términos anteriores conseguiremos mantener un conjunto de variables candidatas a explicar y predecir el modelo que facilitarán la construcción del mismo sin perder capacidad potencial de predicción.

Por último, en esta fase de preanálisis de los datos poblacionales se debe obtener una ***muestra representativa*** de la población objetivo, de forma que los resultados obtenidos en ésta sean aplicables a toda la población. Con ello se consigue además de trabajar con un número de pacientes más reducido, maximizar el número de pacientes en los que se observa el evento cardiovascular en relación con el de los que no se observa el evento. Este planteamiento pretende capturar de un modo más sencillo los distintos perfiles de comportamiento y eliminar posibles sesgos que puedan distorsionar la relación de dependencia capturada por el modelo.

La muestra se obtendrá mediante un muestreo estratificado a partir de la variable estado de *cardiohealth default* que será especificada en el diseño particular de cada estudio. Se aconseja realizar una selección deliberada en la muestra de aquellos pacientes en los que se observa el evento para obtener estimaciones razonablemente precisas y así evitar que al construir el modelo las características de los individuos en los que aparece el evento queden ocultos por la gran proporción de pacientes que no presentan el evento cardiovascular valorado. Habitualmente en modelos de riesgos se considera adecuada una muestra con tamaño entre 10.000 y 20.000 observaciones, con todos los pacientes en los que se ha observado el evento cardiovascular considerado y una extracción aleatoria de al menos, el doble de pacientes de la población en los que no se ha observado el evento con el fin de garantizar la existencia de una muestra de modelización suficientemente grande.

Con el fin de seleccionar el modelo, medir su capacidad de generalización, validar su poder discriminante y calibrarlo, así como comparar los resultados obtenidos por distintas técnicas, siguiendo a HASTIE et al (2009) (61) obtendremos a partir de nuestro conjunto de datos una partición en tres muestras aleatorias simples. Esta partición que presupone una riqueza de datos, estará estratificada por la variable estado *cardiohealth default* indicadora de la aparición o no del evento cardiovascular.

A partir de la muestra seleccionada se obtendrán las tres submuestras siguientes:

1) *Muestra de entrenamiento*, que se utilizará para ajustar el modelo con el criterio de minimizar el error de intervalo. Se le asignará un 40% del total de observaciones de la muestra.
2) *Muestra de validación*, a través de la cual se estimará el error de predicción esperado con el fin de seleccionar el modelo adecuado. Le asignaremos un 30% de los pacientes de la muestra.
3) *Muestra test*, con la que se calculará el error de generalización del modelo finalmente elegido. Se le asigna el 30% restante de la muestra de pacientes considerada.

7.3. EXPLORACIÓN DE LOS DATOS DE ENTRENAMIENTO

Una vez construida y divida la muestra de trabajo en las tres submuestras, de entrenamiento o ajuste, de validación y test, la siguiente fase a acometer consiste en explorar, analizar y preparar los datos de entrenamiento para conseguir modelos de predicción del estado de *cardiohealth default* lo más generalizables y eficientes posible.

La primera tarea en la exploración de la muestra de entrenamiento consiste en hacer una revisión de la distribución de las variables mediante medidas descriptivas, tanto de localización como de escala, gráficas de dispersión, histogramas y funciones de densidad de la variables, etc., para comprobar que no existen diferencias significativas respecto a la distribución en el conjunto de desarrollo completo, de modo que sigan siendo válidas para la muestra de entrenamiento las conclusiones preliminares sobre la capacidad explicativa de las variables independientes.

A continuación, dado que el interés se centra en los modelos logísticos parcialmente lineales, se debe analizar el comportamiento lineal o no de las variables explicativas del riesgo cardiovascular en relación con el *logit* de la probabilidad de *cardiohealth default*. Una vez detectadas aquellas variables que pertenecerán a la componente lineal del modelo, y si ésta no es suficiente para la descripción del modelo de estimación del estado de *cardiohealth default*, se emplearán otros métodos orientados a detectar y confirmar la no linealidad.

7.3.1. Exploración de la linealidad de las variables explicativas del riesgo en relación con el logit de la probabilidad de cardiohealth default

La facilidad de interpretación de los modelos lineales para explicar el estado de *cardiohealth default*, además de su gran capacidad de generalización, nos conducen a que sea la linealidad del modelo y por ello de las variables explicativas, un tema de fundamental importancia.

Además, dado que el interés de la propuesta se centra en los modelos logísticos parcialmente lineales, el aspecto más importante de la exploración de los datos de entrenamiento a efectos de la construcción de modelos de riesgo está relacionado con el comportamiento lineal o no lineal de las variables explicativas sobre la relación de dependencia de éstas con el *logit* de la probabilidad de *cardiohealth default*. Esta fase fundamental de exploración pretende detectar aquellas variables con un comportamiento lineal para formar la componente lineal del modelo logístico y detectar también aquellas que presentan una clara relación no lineal para formar una componente no lineal. Con ambas se conformará una expansión de funciones de base para estimar el *logit* de la probabilidad de *cardiohealth default*.

Se podría plantear la selección automática de variables basada en la regresión logística paso a paso. El problema es que esta técnica adolece de la problemática de la inestabilidad y el sesgo que presentan las estimaciones de los coeficientes de regresión, sus errores estándar y los intervalos de confianza en los casos en los que entre las interrelaciones de las variables exista demasiado *ruido*. Por ello, se descarta esta técnica y se plantea una metodología diferente.

La metodología propuesta para la detección del comportamiento lineal o no de las variables explicativas con el *logit* de la probabilidad de *cardiohealth default a posteriori*, consta de dos fases. En una primera fase se ajustará a los datos de entrenamiento un modelo lineal con todas las variables de la preselección; y en una segunda fase se utilizará una variante *bootsrap* de la regresión logística lineal *backward*, que es una de las técnicas más usuales de selección automática. Analicemos en detalle estas dos fases de la metodología propuesta:

Fase I: Se comienza ajustando los datos de entrenamiento a un modelo logístico lineal con todas las variables de la preselección. De esta forma, además de profundizar en la selección de variables explicativas del riesgo cardiovascular

podremos explorar la estructura formal y particularmente su linealidad. Se ajustarán los datos a un modelo lineal a través de la Regresión Logística Lineal, *LLR*. Una vez comprobada la bondad del ajuste por el estadístico $-2logver$ y los Criterios de Información AIC o BIC, podremos con la ayuda de los test Chi – cuadrado de Wald y razón de verosimilitud, decidir sobre la significación de cada variable en la estructura lineal del modelo. Se construirá un rango de las variables explicativas preseleccionadas sobre los niveles de significación $\alpha = 0{,}01,\ \alpha = 0{,}05\ y\ \alpha = 0{,}10$.

Fase II: En esta fase se emplea una de las técnicas más usuales de selección automática, la regresión logística lineal *backward* pero con la variante de la utilización de muestreo *bootstrap* sobre n ($n \geq 200$) remuestras bootstrap.

Esta técnica, que se conoce con el nombre de ***Regresión Logística Lineal paso a paso Backward Bagged, BBLR_Bag,*** comparte las cualidades especializadas como método de selección automático de la regresión logística lineal *backward* y al mismo tiempo soluciona los problemas de inestabilidad y sesgo que presentan las estimaciones en presencia de *ruido*. Además como método *backward* se caracteriza por comenzar añadiendo todas las variables en el modelo e ir eliminando las menos significativas de forma iterativa. Sobre las n remuestras bootstrap se obtendrán los vectores de estimaciones de los parámetros correspondientes a los n ajustes del modelo, y a continuación se construirá la media de los parámetros. Este análisis se puede implementar mediante el lenguaje SAS® sobre la base de los procedimientos LOGISTIC con métodos de selección *backward* y SURVEYSELECT de SAS® ejecutándose n veces. Se considerarán los niveles de significación $\alpha = 0{,}01,\ \alpha = 0{,}05\ y\ \alpha = 0{,}10$ del test Chi – cuadrado de Wald para detectar y eliminar, en un proceso hacia atrás, las variables menos explicativas del modelo.

Como resultado de la aplicación de *BBLR_Bag* sobre las n ($n \geq 200$) remuestras *bootstrap* se obtendrán n vectores de estimaciones de los parámetros correspondientes a los n ajustes del modelo, y a continuación se construirá la media de los parámetros.

Esta propuesta es más costosa computacionalmente, ya que para cada muestra *bootstrap* se debe considerar la complejidad del proceso iterativo de exclusión de

variables frente al proceso más simple de una regresión lineal sobre tal remuestra, pero puede ser buena herramienta para detectar la no linealidad en presencia de ruido.

Para medir la bondad del ajuste de cada modelo, es decir, las discrepancias entre los datos observados y los datos pronosticados sobre la muestra de entrenamiento, se utilizarán tres estadísticos basados en el estadístico $-2logver$, con la siguiente formulación:

$$-2logver = -2logL\left(Y, \hat{C}(X)\right) = -2\sum_{i=1}^{N} log\left(\hat{P}\left(Y = y_j | X = x_i\right)\right)$$

1. El *pseudo – coeficiente de determinación* de NAGELKERKE:

$$\widetilde{R}^{\,2} = \frac{1 - \left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}}}{1 - \left(verosim(M_0)\right)^{\frac{2}{N}}}$$

donde M representa el modelo con todas las variables explicativas consideradas y M_0 es el modelo con sólo el término intercepto, llamado *modelo nulo*.

Si el modelo ajustado predice perfectamente los resultados y tiene verosimilitud 1, el $\text{pseudo} - \text{R}^2$ de Nagelkerke es $\widetilde{\text{R}}^{\,2} = 1$. Sin embargo, si el modelo completo no mejora al modelo consistente en sólo el coeficiente intercepto, el coeficiente $\widetilde{\text{R}}^{\,2} > 0$, por lo que el rango total $[0, 1]$ de los mínimos cuadrados ordinarios no está cubierto.

El *pseudo – coeficiente de determinación* $\widetilde{R}^{\,2}$ no puede ser interpretado de forma independiente o comparando distintos conjuntos de datos. Su validez y utilidad se centra en la evaluación de varios modelos de predicción sobre la misma muestra de datos con la misma variable respuesta. En esta situación, un $\widetilde{R}^{\,2}$ más alto indica que el modelo predice mejor la respuesta.

Este test de bondad de ajuste por estar basado en $-2logver$, conduce con frecuencia a rechazar modelos aceptables a la vez que a aceptar algunos que resultan menos parsimoniosos de lo que debieran, lo que le restaría capacidad de generalización y facilidad explicativa requerida al modelo. Por tanto, se proponen otros criterios para medir la bondad del ajuste, basados en $-2logver$

penalizada, el Criterio de Información de *AKAIKE* (AIC), y el Criterio de Información Bayesiano de *SCHWARZ* (BIC). En estas medidas el término de penalización es el encargado de corregir la complejidad del modelo o, en otros términos, el sobreajuste.

2. El Criterio de Información de *AKAIKE*:

$$AIC = -2logver + 2p$$

siendo p el número de variables explicativas de riesgo cardiovascular de entrada.

AIC es asintótico por lo que se requieren muestras grandes. Además, el número máximo de parámetros no puede exceder $2pN$, donde N es el número de observaciones.

3. El Criterio de Información Bayesiano:

$$BIC = -2logver + log(N)p$$

Para AIC y BIC cuanto menor es su valor, mejor es el ajuste. Dos de las mayores fortalezas de estas dos medidas son:

- Se puede comparar el ajuste de diferentes modelos, incluso cuando los modelos no están anidados. La idea básica es comparar la verosimilitud relativa de los dos modelos en vez de analizar la desviación absoluta de los datos observados de un modelo particular.
- Como medidas de información que son, penalizan la inclusión de variables que no mejoran significativamente el ajuste. En particular, con grandes muestras las medidas de información pueden conducir a modelos más parsimoniosos.

Con el fin de facilitar la toma de decisión sobre qué variables se considerarán lineales para formar parte de la componente lineal del modelo, se elaborará una tabla de resultados que recoja los parámetros estimados para las variables explicativas y rango de las variables explicativas ordenadas por el número de veces que cada variable alcanza el grado de significación $\alpha = 0{,}01$, $\alpha = 0{,}05$ y $\alpha = 0{,}10$. Igualmente se tabularán los estadísticos de bondad de ajuste para el modelo obtenido mediante

Regresión Logística Lineal y el obtenido mediante Regresión Logística Lineal paso a paso *Backward Bagged*.

7.3.2. Exploración de la estructura de la distribución de las variables con linealidad no significativa

Antes de explorar la no linealidad de las variables que las técnicas *LLR* y *BBLR_Bag* no detectaron con linealidad significativa, se analizará en detalle si presentan o no estructura continua. Para ello se propone la representación de las distribuciones de las variables utilizando la técnica de estimación de la densidad por núcleos Gaussianos. Este hecho es importante para la configuración de la componente no lineal del modelo, ya que las componentes no paramétricas requieren variación continua.

Una vez que la metodología de exploración de la linealidad de las variables explicativas facilita la posible estructura del modelo en su componente lineal y si ésta no es suficiente para la descripción del modelo de estimación del estado de *cardiohealth default* de los pacientes, se emplearán otros métodos orientados a detectar y confirmar la no linealidad. Entre los métodos de detección de la no linealidad destacar el más sencillo y conocido que consiste en la utilización de Gráficos Exploratorios de Dispersión, otro clásico es el Test de Box – Tidwell o el Método de los Residuos Acumulados, que consiste en una técnica de chequeo de modelos lineales generalizados basado en la suma acumulada de residuos complementada con el test del supremo de Kolmogorov.

7.4. ESPECIFICACIÓN Y AJUSTE DEL MODELO

7.4.1. Introducción

La especificación de la estructura funcional es una de las cuestiones fundamentales en la construcción de un modelo estadístico, ya que una especificación incorrecta puede dar lugar a estimaciones sesgadas o a coeficientes ineficientes.

Para valorar la estructura funcional especificada, una de las primeras cuestiones que se deben plantear es si la forma funcional del modelo es correcta, y otra si todas las variables explicativas pertinentes están incluidas en el modelo y no se ha incluido ninguna de las irrelevantes.

La idoneidad del modelo dependerá del grado de conocimiento que se tenga tanto de la distribución conjunta de la variable respuesta y las variables explicativas, como de las distribuciones conjuntas de las variables explicativas condicionadas a la variable estado de *cardiohealth default* y *no cardiohealth default.* Además deberá obtenerse en función de los datos disponibles, que no siempre son los más adecuados para especificar los mejores modelos.

Una buena formulación de un modelo deberá satisfacer las siguientes condiciones:

a) El modelo explica de forma adecuada el comportamiento del paciente frente al *cardiohealth default*.
b) El modelo predice de forma significativamente correcta la probabilidad de *cardiohealth default*.
c) El modelo clasifica correctamente a nuevos pacientes distintos a los utilizados en su entrenamiento.
d) La puntuación otorgada por el modelo de calificación a un paciente es fácilmente interpretable en función del *peso* de cada variable en el modelo.

La combinación de estas cuatro condiciones determina las propiedades que ha de tener el modelo más idóneo. Se buscará un equilibrio entre las tres primeras características que se presuponen en un modelo óptimo desde el punto de vista estadístico y la facilidad de interpretación de la función de riesgo. Esta última característica se justifica desde el punto de vista médico y de aplicación del sistema de cuantificación en prevención primaria, con el fin de poder transmitir al paciente su nivel de riesgo cardiovascular de una forma clara y así conseguir la motivación del mismo sobre el control de los factores de riesgo, principalmente de aquellos que provengan de conductas modificables.

7.4.2. Conceptos Generales

Como hemos visto en el apartado 6.1, la relación de dependencia entre $g\big(P(Y = 1|X = x)\big)$ y las variables explicativas X que se quiere estimar adopta la forma de una función $C(X)$, *función de calificación* de pacientes, que se asume es miembro de una familia de funciones $\boldsymbol{F} = \big\{C: x \in \mathbb{R}^p \longrightarrow g\big(P(Y = 1|X = x)\big) \in \mathbb{R}\big\}$. Para ajustar el modelo, una vez especificadas la función de enlace $g(\cdot)$ y $C(\cdot)$, a partir de una muestra

de observaciones de la población $\boldsymbol{\tau} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1,\dots,N}$ se obtendrá un estimador $\hat{C}(\cdot)$ óptimo a través de un criterio de optimización fijado.

Según se ha justificado en el apartado 6.2.1, dado que la variable respuesta Y indicadora del estado de *cardiohealth default* es binaria, la función de enlace que se considerará es la transformación logística.

Para decidir entre varias funciones posibles cuál describe mejor la dependencia observada se introducen los conceptos de función de pérdida, riesgo esperado y riesgo empírico.

Definición 7.1: Sea $X \in \mathbb{R}^p$ vector aleatorio de las variables explicativas e $Y \in \mathbb{R}$ una variable aleatoria perteneciente a una familia de variables respuesta Y, con distribución conjunta $P(X, Y)$. Sea $C(X)$ una función de la familia $\boldsymbol{F}$ de funciones para predecir Y dados los valores de X. Se dice que la función $\ell(Y, C(X))$ es una *función de pérdida* para penalizar los errores en la predicción si es acotada y mide el coste de la discrepancia entre la función pronóstico $C(X)$ y la variable aleatoria respuesta Y.

La función de pérdida $\ell(Y, C(X))$ es una herramienta clave para estimar el modelo, puesto que a partir de ella se obtiene el criterio para ajustarlo a los datos y dado que en este caso se usa en problemas de optimización, debe ser convexa.

Definición 7.2: Se llama *riesgo esperado* asociado a la función de pérdida $\ell(Y, C(X))$ a la cantidad:

$$R_\ell(Y, C(X)) = E_P[\ell(Y, C(X))] \qquad (7.1)$$

donde E_P es la esperanza matemática con respecto a la distribución conjunta $P(X, Y)$ del vector aleatorio de variables explicativas X y la variable aleatoria respuesta Y.

Un criterio razonable para elegir $C(X)$ consiste en minimizar el riesgo esperado (6.16), y dado que $P(X, Y) = P(Y|X)P(X)$, se tiene que:

$$E_P[\ell(Y, C(X))] = E_X E_{Y|X}[\ell(Y, C(X))|X] \qquad (7.2)$$

por lo que es suficiente minimizar punto a punto (7.2):

$$C(x) = \min_C E_{Y|X}[\ell(Y, C(X))|x] \qquad (7.3)$$

Definición 7.3: Dada una muestra aleatoria $\{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1,\dots,N}$ y siendo $\ell(Y, C(X))$ una función de pérdida para la predicción de Y por $C(X)$,

a) Se llama *función de pérdida empírica* para la observación (x_i, y_i) a la cantidad $\ell(y_i, C(x_i))$.

b) La expresión

$$\hat{R}_{emp}(Y, C(X)) = \frac{1}{N}\sum_{i=1}^{N} \ell(y_i, C(x_i))$$

es la pérdida media empírica, estimador del riesgo esperado (7.1), llamado *Riesgo Empírico*. En los problemas de optimización es equivalente optimizar el riesgo empírico que optimizar la cantidad $\hat{L}_{emp}(Y, C(X)) = \sum_{i=1}^{N} \ell(y_i, C(x_i))$ a la que llamaremos *pérdida empírica*.

Un ejemplo de función de pérdida particularmente importante este trabajo lo constituye la **función de pérdida logística**, y que tiene la siguiente expresión:

$$\ell(Y, C(X)) = -\left[YC(X) - log\left(1 + exp\big(C(X)\big)\right)\right]$$

Se llama *estimador logístico de la función de calificación* $C(X)$ a la función $\hat{C}(X) = logit\left(\hat{P}(Y = 1|X = x)\right)$, si es solución del problema de optimización siguiente:

$$\underset{C}{Min}\left\{-\sum_{i=1}^{N}\left[y_i C(x_i) - log\left(1 + exp\big(C(x_i)\big)\right)\right]\right\} \qquad (7.4)$$

conocido como *Problema de Optimización Logístico*.

Una vez obtenida la función de calificación estimada $\hat{C}(X)$, la *probabilidad de cardiohealth default a posteriori* se obtendrá a través del estimador:

$$\hat{P}(Y = 1|X = x) = \Lambda\big(\hat{C}(X)\big) \qquad (7.5)$$

El planteamiento del problema de optimización se puede llevar a cabo mediante estimadores regularizados con el fin de evitar el problema del sobre ajuste o infra ajuste del modelo, lo que le restaría capacidad de generalización.

La generalización se consigue principalmente "suavizando el modelo", es decir, consiguiendo modelos de tendencia antes que modelos muy ajustados localmente. Esto se consigue a través de una funcional de suavizado $J\big(C(X)\big)$, también llamada de penalización o regularización, de modo que bajos valores de la funcional corresponden a funciones suavizadas.

Definición 7.4: Para una muestra aleatoria $\boldsymbol{\tau} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1,\dots,N}$, el problema de optimización

$$\underset{C(X)}{Min} \sum_{i=1}^{N} \ell\big(y_i, C(x_i)\big) + \lambda J\big(C(X)\big) \qquad (7.6)$$

se conoce como *Problema de Minimización de la Pérdida Empírica Regularizada*.

7.4.3. Estimación del *SPANISH CARDIOVASCULAR RISK SCORECARD*

Comenzamos por establecer una serie de hipótesis sobre las que plantearemos la estructura funcional del modelo y una metodología apropiada para su estimación. Estas hipótesis están inspiradas en los Modelos de Probabilidad Generalizados, en el carácter binario de la variable respuesta, así como en las expansiones lineales por funciones de base y en el hecho de que para estimar los modelos de riesgo se cuenta casi siempre con información limitada e imperfecta, lo que conduce al principio de inducción.

Hipótesis 1: El modelo expresa la relación existente entre la transformación logística de la probabilidad de *cardiohealth default*, $\Lambda(\cdot)$ con las propiedades necesarias para asegurar que $P(\cdot) = \Lambda\big(C(X)\big)$ es una probabilidad, y la *función de calificación* (función de las variables de riesgo explicativas del estado de *cardiohealth default*):

$$logit\big(P(Y = 1|X)\big) = C(X)$$

Con esta hipótesis se pretende que nuestro modelo pertenezca a la familia de los Modelos de Probabilidad de Generalizados, puesto que uno de nuestros objetivos consiste en que el modelo nos proporcione la *probabilidad de cardiohealth default*.

Hipótesis 2: La *función de calificación* $C(X)$, es una expansión lineal de funciones de base del vector de variables explicativas X según (6.23).

El objetivo de esta hipótesis consiste en construir modelos más flexibles, que contemplen la no linealidad aumentando o reemplazando el vector de variables explicativas originales con variables adicionales, pudiendo ser éstas transformaciones de X, tales como pesos de la evidencia, polinomiales, splines de regresión cúbicos restringidos, funciones constantes a trozos resultado de particiones recursivas, funciones lineales a trozos, funciones bisagra MARS, funciones de base radial Gaussiana, etc. y usar modelos lineales en este nuevo espacio expansionado de las variables de entrada resultantes.

En presencia de la no linealidad de las variables explicativas esta hipótesis es clave y adquiere su significado, puesto que el hecho de que la *función de calificación* $C(X)$ sea una expansión lineal de funciones de base de las variables explicativas del riesgo permite mantener la estructura formal de un modelo interpretable.

Hipótesis 3: Para ajustar el modelo, consideraremos como hipótesis general que los datos son finitos e imperfectos, y la información que nos proporcionan es limitada, por lo que el Principio de Inducción constituye un método adecuado de estimación del modelo.

Una vez fijadas las hipótesis estamos en disposición de fijar la estructura funcional del modelo y la metodología para su estimación.

La estructura formal del modelo que se propone para la estimación del riesgo cardiovascular, según se ha justificado en el apartado 6.4, es el denominado ***Modelo Logístico Lineal Híbrido (HLLM)***, que es un modelo logístico lineal por expansiones lineales híbridas de funciones de base.

La estructura formal del ***HLLM*** según (6.25) se expresa de la forma:

$$logit\big(P(Y=1|X=x)\big) = \beta_0 + \sum_{r=1}^{p_1} \beta_r U_r + \sum_{r=1}^{p_2} \left(\sum_{k=1}^{K_r} \theta_{r,k}\, h_{r,k}(V_r) \right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{\theta}^T \boldsymbol{H}(\boldsymbol{V}) \quad (7.7)$$

donde $p_1 + \sum_{r=1}^{p_2} K_r = q$ y $p_1 + p_2 = p$, $\boldsymbol{X}^T = (\boldsymbol{U}^T, \boldsymbol{V}^T)$, con $\boldsymbol{U}^T = (U_1, \dots, U_{p_1})$ vector de variables lineales y $\boldsymbol{V}^T = (V_1, \dots, V_{p_2})$ vector de variables no lineales, $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_{p_1})$, $\boldsymbol{H}(\boldsymbol{V}) = \left(\boldsymbol{H}_1(V_1), \dots, \boldsymbol{H}_{p_2}(V_{p_2}) \right)$, $\boldsymbol{H}_r(V_r) = \left(h_{r,1}(V_r), \dots, H_{r,K_r}(V_r) \right)$ con $r = 1, \dots, p_2$, $\boldsymbol{\theta}^T = (\theta_1, \dots, \theta_{p_2})$ y $\boldsymbol{\theta}_r^T = (\theta_{r,1}, \dots, \theta_{r,K_r})$.

La función objetivo asociada a la pérdida logística empírica regularizada para el modelo anterior viene dada por:

$$L_{emp\ \ell_\Lambda}(Y, \beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{\theta}^T \boldsymbol{H}(\boldsymbol{V})\ ) =$$

$$= -\left[Y^T \big(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{\theta}^T \boldsymbol{H}(\boldsymbol{V})\big) - \mathbf{1}^T log\left(1 + exp\big(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{\theta}^T \boldsymbol{H}(\boldsymbol{V})\big)\right)\right]$$
$$+ \lambda\, J\big(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{U} + \mathbf{1}^T \boldsymbol{\theta}^T \boldsymbol{H}(\boldsymbol{V})\big)$$

Por lo que el modelo se estima resolviendo el *Problema de Optimización Logístico* siguiente:

$$\underset{\beta_0,\boldsymbol{\beta},\boldsymbol{\theta}}{Min}\ L_{emp\ \ell_\Lambda}(Y, \beta_0 + \boldsymbol{\beta}^T\boldsymbol{U} + \mathbf{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\ ) \qquad (7.8)$$

En caso de ser $\lambda > 0$ estaremos ante el caso regularizado, en el que habría que especificar $J\big(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{U} + \mathbf{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\big)$ como una función convexa y tal que existan su primera y segunda derivada respecto de los vectores de parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$.

Resolver el problema de optimización (7.8) equivale a resolver el sistema de ecuaciones normales que se obtiene al igualar el vector gradiente a cero. El sistema de ecuaciones normales representa un sistema de $q + 1$ ecuaciones no lineales con $q + 1$ incógnitas $(\beta_0, \beta_1, \dots, \beta_q)$, que se resolverá mediante el algoritmo de Newton –Raphson, que además del vector gradiente requiere la matriz Hessiana.

Por tanto, para resolver el problema de optimización (7.8) es necesario establecer las dos hipótesis siguientes:

i. $J\big(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{U} + \mathbf{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\big)$ es una función convexa.

ii. $\exists\ J'(\boldsymbol{\beta},\boldsymbol{\theta})$ y $J''(\boldsymbol{\beta},\boldsymbol{\theta})$.

La solución del problema de optimización (7.8) bajo los supuestos anteriores se expresa como:

$$(\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{nuevo} = \Big[\big(\mathbf{1},\boldsymbol{U},\boldsymbol{H}(\boldsymbol{V})\big)^T W\big(\mathbf{1},\boldsymbol{U},\boldsymbol{H}(\boldsymbol{V})\big) + \lambda\big(0, J''(\boldsymbol{\beta},\boldsymbol{\theta})\big)\Big]^{-1}\big(\mathbf{1},\boldsymbol{U},\boldsymbol{H}(\boldsymbol{V})\big)^T\boldsymbol{W}\boldsymbol{z} \qquad (7.9)$$

donde,

$$\boldsymbol{z} = ((\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{anterior})^T\big(\mathbf{1},\boldsymbol{U},\boldsymbol{H}(\boldsymbol{V})\big) + \boldsymbol{W}^{-1}(y - P) + \lambda\Big(\big(\mathbf{1},\boldsymbol{U},\boldsymbol{H}(\boldsymbol{V})\big)^T\boldsymbol{W}\Big)^{-1}\big[\big(0\ , J''(\boldsymbol{\beta},\boldsymbol{\theta})\big)(\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{anterior} - \big(0, J'(\boldsymbol{\beta},\boldsymbol{\theta})\big)\big] \qquad (7.10)$$

es la variable respuesta ajustada o variable respuesta de trabajo.

La nueva expansión lineal de funciones de base se expresa según la siguiente igualdad:

$$((\beta_0,\boldsymbol{\beta},\boldsymbol{\theta})^{nuevo})^T\big(\mathbf{1},\boldsymbol{U},\boldsymbol{H}(\boldsymbol{V})\big) = \beta_0^{nuevo} + (\boldsymbol{\beta}^{nuevo})^T\boldsymbol{U} + (\boldsymbol{\theta}^{nuevo})^T\boldsymbol{H}(\boldsymbol{V})$$

La complejidad del modelo está controlada por el número de variables lineales y por la expansión de funciones de base que fijemos.

Además, si lo que se busca es la simplicidad del modelo cabe la posibilidad de considerar el modelo sin regularizar, es decir aquel en el que $\lambda = 0$, lo que implica que tango el modelo como el método de estimación se simplifican bastante, ya que el problema de optimización general de los modelos logísticos expansión lineal de funciones de base se reduce a:

$$\underset{\beta_0,\boldsymbol{\beta},\boldsymbol{\theta}}{Min} -\left[Y^T\left(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{U} + \mathbf{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right) - \mathbf{1}^T log\left(1 + exp\left(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{U} + \mathbf{1}^T\boldsymbol{\theta}^T\boldsymbol{H}(\boldsymbol{V})\right)\right)\right] \quad (7.11)$$

y la solución viene dada por:

$$(\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta})^{nuevo} = \left[\left(\mathbf{1}, \boldsymbol{U}, \boldsymbol{H}(\boldsymbol{V})\right)^T \boldsymbol{W} \left(\mathbf{1}, \boldsymbol{U}, H(\boldsymbol{V})\right)\right]^{-1} \left(\mathbf{1}, \boldsymbol{U}, \boldsymbol{H}(\boldsymbol{V})\right)^T \boldsymbol{W}\boldsymbol{z} \quad (7.12)$$

donde,

$$\boldsymbol{z} = \left((\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta})^{anterior}\right)^T \left(\mathbf{1}, \boldsymbol{U}, \boldsymbol{H}(\boldsymbol{V})\right) + \boldsymbol{W}^{-1}(y - P) \quad (7.13)$$

es la variable respuesta ajustada o variable respuesta de trabajo para el caso no regularizado del problema.

7.4.4. Selección de las Funciones de Base para la componente no lineal del modelo

En general, para la construcción del modelo de riesgo cardiovascular se contará con p_1 variables con influencia lineal sobre la variable estado de *cardiohealth default*, $U_1, \dots, U_{p_1}$ y p_2 variables $V_1, \dots, V_{p_2}$, con influencia desconocida y no lineal.

Con este hecho como hipótesis de partida, el modelo logístico parcialmente lineal y semiparamétrico, adopta la forma estructural dada por (6.20). Nuestro objetivo consiste en convertir dicha estructura en la correspondiente al modelo (6.25), modelo logístico lineal paramétrico, a través de expansiones lineales de funciones de base $\sum_{k=1}^{K_r} \theta_{r,k}\, h_{r,k}(V_r)$. La estructura (6.25) estará perfectamente especificada si para cada variable con influencia no lineal $V_r, r = 1, \dots, p_2$, hayan sido seleccionadas las funciones de base $h_{r,k}(V_r)$, $k = 1, \dots, K_r$.

Para la selección de estas funciones de base se propone la metodología que plantea Mallo Fernández(53) adaptada a nuestro contexto de riesgo cardiovascular. Ésta consta de los tres pasos siguientes:

1. En primer lugar se considera un modelo de partida $M_{inicial}$, modelo logístico lineal

$$logit\big(P(Y=1|X=x)\big)=\beta_0+\sum_{r=1}^{p_1}\beta_r U_r$$

con todas las variables explicativas linealmente significativas que se consideran de interés en riesgo cardiovascular, en la que la función de calificación de pacientes es una expansión lineal de funciones identidad. Este modelo se debe plantear necesariamente, ya que difícilmente cualquier otra estructura resulta más fácilmente interpretable que la lineal.

2. En segundo lugar se hace uso de un método constructivo para seleccionar la combinación lineal de funciones de base "más prometedora" para cada variable no lineal de un conjunto de candidatas e incorporarla al modelo. Este método consiste en ir añadiendo, en cada paso, al modelo inicial las combinaciones lineales cuyas funciones de base resulten significativas, utilizando el test chi – cuadrado, en el modelo logístico conseguido en el paso anterior. En cada paso se irá analizando si la incorporación de funciones de base incide en la significación del modelo de otras variables ya incorporadas, a la vez que se comprobará el ajuste del modelo a los datos de entrenamiento, a través de los pseudo – coeficientes de determinación, McFadden y Nagelkerke, del Error Empírico y de los criterios de Información de Akaike, AIC, y Schwarz, BIC, y el poder discriminante del modelo, a través del área bajo la curva ROC, AUC.

Antes de proceder a la selección de las funciones de base más prometedoras en cada paso, se realiza un proceso de evaluación sobre la muestra de validación, obtenida de forma aleatoria e independiente de la muestra de entrenamiento para tal fin. El proceso sobre la muestra de validación conlleva prácticamente los mismos estadísticos que para la muestra de entrenamiento, que serían: la evaluación del Ajuste del modelo, el Error de validación, (error empírico sobre la muestra de validación), los Criterios de Información AIC y BIC, la Validación del Poder Discriminante del nuevo modelo sobre la muestra de validación, AUC, y la Tasa de Clasificación Errónea.

Se procede secuencialmente según el proceso anterior para todas y cada una de las variables de la componente no lineal. A la vez se construirán modelos alternativos considerando expansiones lineales por funciones base, que aunque

no sean las más prometedoras parezcan en principio adecuadas, o al menos, de interés a efectos comparativos.

3. Como resultado del proceso de construcción expuesto en el punto anterior se llega a un conjunto de modelos con estructura inicialmente válida pero que podría, en principio, sobre ajustar los datos, no poseer la cualidad de facilidad interpretativa, etc. Para evitar tales inconvenientes, en una tercera fase se procede a aplicar técnicas de poda o regularización para reducir el número de funciones de base.

Antes de comenzar a construir el modelo, se fijarán las funciones de base que pueden ser apropiadas para construir las expansiones lineales para cada una de las variables sobre las que no se observa significación en el modelo que hemos denominado $M_{inicial}$, que se corresponde con el modelo logístico lineal (6.17).

Las funciones de base que se podrían considerar, van desde funciones muy sencillas tales como potencias, logaritmos, interacciones entre las variables, hasta otras más complejas como las Funciones de base polinómicas de orden p, Funciones Constantes a Trozos obtenidas a partir de Indicadores de Particiones Recursivas, Pesos de la Evidencia asignados a Tramados de las Variables o a Particiones Recursivas, Splines Cúbicos Restringidos de Stone y Koo, Funciones Bisagra obtenidas por MARS Univariante, Funciones de Base Radial, etc.

La no linealidad de las variables puede presentarse en infinitas formas y es de esperar que haya otras tantas familias de funciones de base capaces de aproximarla expandiendo las variables de riesgo originales a espacios agrandados de Hilbert. El conjunto de todas las familias de funciones base susceptibles de ser seleccionadas para la construcción de un Modelo Logístico Lineal por expansiones lineales Híbridas de funciones de base componen lo que se denomina ***Diccionario de Funciones de Base***. La investigación al respecto de las funciones de base óptimas en cada contexto y la elaboración de un detallado *Diccionario de Funciones de Base* acompañado de algún método para controlar la complejidad del modelo, supone un campo de exploración abierto que conformaría una valiosa herramienta de aplicación real en la estimación del riesgo en las distintas áreas.

La propuesta que se hace en esta Tesis se basa en tres de los tipos de funciones de base sobre las que ya se ha demostrado su buen funcionamiento en la estimación del riesgo. Las funciones de base a las que nos referiremos son las Funciones Constantes a Trozos obtenidas a partir de Indicadores de Particiones Recursivas, los Pesos de la Evidencia asignados a Tramados de las Variables o a Particiones Recursivas y los Splines Cúbicos Restringidos de Stone y Koo.

7.4.4.1. Funciones Constantes a Trozos obtenidas a partir de Indicadores de Particiones Recursivas

Esta técnica consiste en dividir el espacio de características en un conjunto de híper – rectángulos, que constituyen una Partición Recursiva para así pasar a ajustar un simple modelo igual a una constante en cada uno.

Troceando apropiadamente el rango de la variable X_j en regiones disjuntas se obtiene una partición del rango de V_r en K_j regiones disjuntas $\{R_{r1}, \dots, R_{rK_j}\}$, donde $Rango(V_r) = \cup_{k=1}^{K_j} R_{rk}$ con $R_{rk} \cap R_{rl} = \emptyset$ para todo $k \neq l$. Definiendo las funciones de base como indicadores de cada una de las regiones de la partición del rango de X_j resulta un modelo con contribuciones constantes a trozos para la variable.

$$h_r(X_j) = I_{[X_j \in R_{jk}]}\ , \quad r = 1, \dots, q, \ \ j = 1, \dots, p$$

Según el método utilizado para obtener la partición de las regiones que componen el rango de la variable tendremos diferentes funciones pero siempre constantes a trozos. Las particiones más habituales en la industria de los sistemas de calificación del riesgo son las particiones estadísticas automáticas óptimas, aquellas obtenidas de forma automática con criterios estadísticos únicamente, las particiones estadísticas con criterios de riesgos (SIDDIQI, 2006) (62) y las particiones recursivas binarias conseguidas por el algoritmo CART de los árboles de decisión y de clasificación, TREE, (BREIMAN, et al., 1984) (63). En nuestro caso el método que seleccionaremos será el **algoritmo CART de los árboles de decisión**.

Los árboles de decisión son conceptualmente simples y muy atractivos ya que están dotados de una gran facilidad de interpretación. Entre sus cualidades destaca el hecho de que las ramas del árbol simulan bastante bien el proceso humano para la toma de

decisiones a la vez que definen directamente las reglas de asignación, por lo que sus resultados son operativos inmediatamente. Una de sus mayores fortalezas es que detectan de forma automática estructuras complejas entre variables. Por otra parte, junto al hecho de que minimizan el preanálisis de los datos, puesto que tienen una gran capacidad para trabajar con un nivel de ruido relativamente alto y con datos faltantes, son computacionalmente muy eficientes. Por otro lado, las debilidades no son pocas ya que presentan una alta varianza y la probabilidad estimada no se encuentra entre las mejores ya que es considerada como función constante a trozos, no siendo ésta la forma usual de la función subyacente real. Por ello, la aplicación en nuestro modelo de esta técnica se hará con una gran dosis de cautela.

7.4.4.2. Pesos de la Evidencia asignados a Tramados de las Variables o a Particiones Recursivas

Esta familia de funciones de base se obtiene definiendo para la variable no lineal original V_r una función de base en la forma:

$$h_r(V_r) = \sum_{k=1}^{K_r} I_{[V_r \in R_{rk}]} \, WOE(R_{rk})$$

donde $\{R_{r1}, \dots, R_{rK_r}\}$ es una partición del rango de V_r en K_r regiones disjuntas, $Rango(V_r) = \cup_{k=1}^{K_r} R_{rk}$ con $R_{rk} \cap R_{rl} = \emptyset$ para todo $k \neq l$ y siendo $WOE(R_{rk}) = ln\left(\frac{P_{Buenos\ en\ R_{rk}}}{P_{Malos\ en\ R_{rk}}}\right)$ el peso de la evidencia del *cardiohealth default* para la subregión R_{rk} del rango de la variable V_r.

$P_{Buenos\ en\ R_{rk}} = \frac{Número\ de\ pacientes\ Buenos\ en\ R_{rk}}{Número\ total\ de\ pacientes\ Buenos}$, es la proporción de pacientes buenos en la región R_{rk} (entendiendo como paciente bueno aquel en el que no se ha observado la presencia de un evento cardiovascular).

$P_{Malos\ en\ R_{rk}} = \frac{Número\ de\ pacientes\ Malos\ en\ R_{rk}}{Número\ total\ de\ pacientes\ Buenos}$, es la proporción de pacientes malos en la región R_{rk} (entendiendo como paciente malo aquel en el que se ha observado la presencia de un evento cardiovascular).

Es decir, cada función de base $h_r(X)$ se define como un indicador para el WOE de cada una de las subregiones R_{rk} del rango de V_r. Sustituyendo la variable V_r en el modelo por la variable $\sum_{k=1}^{K_r} I_{[V_r \in R_{rk}]}\, WOE(R_{rk})$, resulta un modelo con contribuciones constantes a trozos para la variable V_r, siendo la constante en cada región o tramo de la variable el peso de la evidencia (***Weigth Of Evidence***) del *cardiohealth default* para la variable en esa región, a x_{ir} se le asigna $\sum_{k=1}^{K_r} I_{[x_{ir} \in R_{rk}]}\, WOE(R_{rk})$.

Dado que la partición del rango de V_r es disjunta, x_{ir} sólo puede pertenecer a una de las subregiones R_{rk} , en la que $I_{[x_{ir} \in R_{rk}]}$ vale 1 y vale 0 en las restantes subregiones, por lo que el valor de la variableV_r sobre el paciente $i-ésimo$ se sustituye por el peso de la evidencia de la subregión del rango de V_r a la que pertenece el paciente.

Asignar el peso de la evidencia a los pacientes que integran una región del rango de la variable parece muy razonable, por cuanto esa cantidad integra toda la información que sobre el *cardiohealth default* y *no cardiohealth default* aporta la pertenencia de un paciente a una región del rango de la variable.

Las funciones de este tipo tienen la ventaja de la facilidad de explicación a los pacientes, puesto que la puntuación del mismo en una variable de este tipo es directamente proporcional al peso del *cardiohealth default* frente al *no cardiohealth default* en la región en la que el paciente se sitúa para la variable en cuestión.

7.4.4.3. Splines Cúbicos Restringidos de Stone y Koo

El término *spline* hace referencia a una amplia clase de funciones que son utilizadas en aplicaciones que requieren la interpolación de datos, o un suavizado de curvas. Las funciones para la interpolación por *splines* normalmente se determinan como minimizadores de la rugosidad sometidas a una serie de restricciones.

Se dice que una polinomial troceada $h(x) = \{h_s\}_{s=1}^{q-1}$ con q nudos, para los que $\xi_1 < \xi_2 < \cdots < \xi_q$, es un ***spline de orden*** $\boldsymbol{M \geq 0}$ si satisface las siguientes condiciones:

i. En cada intervalo $[\xi_{i-1}, \xi_i)$, $h(x)$ es un polinomio de grado menor o igual que $M-1$.

ii. $h(x)$ tiene derivada de orden $M-2$ continua en $[\xi_1, \xi_q)$

$$h(x) = \begin{cases} h_1(x) = a_1 + b_1 x + \cdots + d_1 x^{M-1} & x \in [\xi_1, \xi_2) \\ h_2(x) = a_2 + b_2 x + \cdots + d_2 x^{M-1} & x \in [\xi_2, \xi_3) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ h_{q-1}(x) = a_{q-1} + b_{q-1} x + \cdots + d_{q-1} x^{M-1} & x \in [\xi_{q-1}, \xi_q) \end{cases}$$

Los polinomios h_{s-1} y h_s interpolan el mismo valor en el punto ξ_s para $s = 2, \ldots, q-1$

Los ***splines de regresión*** son aquellos en los que implícitamente se asume que el orden y el número de nodos no están fijados a priori, al igual que su ubicación. Los *splines* de regresión más utilizados son los de orden $M = 4$ y se les conoce como *splines cúbicos*.

El problema más importante que presenta el ajuste por *splines de regresión* es la elección del conjunto de nudos, lo que influye notablemente en los resultados del ajuste. Si el número de nudos es cero el ajuste es lineal, mientras que valores del número de nudos próximos al tamaño muestral conducen al sobreajuste. Este problema puede evitarse con los ***Splines Cúbicos Restringidos*** **de Stone y Koo** (64), que son un caso particular de los *splines de regresión* con magníficas propiedades teóricas y prácticas.

La idea de Stone y Koo consiste en ajustar un efecto principal continuo por una segmentación de suavizadores polinomiales de grado 3, para lo que se generan fórmulas para funciones base que permiten ajustar *splines* restringidos a la linealidad en las colas, es decir, linealidad por encima del último nudo y por debajo del primero.

Los *splines* $k - anudados$ de Stone y Koo, además de ser conservadores comparados con las polinomiales puras en el sentido de que la extrapolación fuera del rango de los datos es una línea recta en vez de una curva polinomial, se construyen con un número relativamente pequeño de funciones de base de cálculo sencillo. Esto conduce a modelos menos complejos y más parsimoniosos, por lo que se están revelando como muy eficaces en campos como la Medicina.

Si V_r es una variable continua que forma parte de la componente no lineal del modelo y q es el número de nudos introducidos sobre el eje V_r, localizados en $\xi_1 < \xi_2 < \cdots < \xi_q$, las funciones de base generadas por el método de Stone y Koo para V_r, $h_1(V_r)$, …, $h_{q-1}(V_r)$, se obtienen de las siguiente forma:

$h_1(V_r) = V_r$

$$h_s(V_r) = \left(V_r - \xi_{r(s-1)}\right)_+^3 - \frac{\left(V_r - \xi_{r(q-1)}\right)_+^3 \left(\xi_{rq} - \xi_{r(s-1)}\right)}{\xi_{rq} - \xi_{r(q-1)}} + \frac{\left(V_r - \xi_{rq}\right)_+^3 \left(\xi_{r(q-1)} - \xi_{r(s-1)}\right)}{\xi_{rq} - \xi_{r(q-1)}}$$

para $s = 2, \dots, q-1$ y donde $(Z)_+ = \begin{cases} Z & si\ Z > 0 \\ 0 & si\ Z \leq 0 \end{cases}$ y la expansión de V_r combinación lineal de tales funciones de base se expresa entonces de la forma:

$$RCS\_V_r = \beta_r V_r + \sum_{s=2}^{q-1} \theta_{r+s-1} \left(\left(V_r - \xi_{r(s-1)}\right)_+^3 - \frac{\left(V_r - \xi_{r(q-1)}\right)_+^3 \left(\xi_{rq} - \xi_{r(s-1)}\right)}{\xi_{rq} - \xi_{r(q-1)}} + \frac{\left(V_r - \xi_{rq}\right)_+^3 \left(\xi_{r(q-1)} - \xi_{r(s-1)}\right)}{\xi_{rq} - \xi_{r(q-1)}} \right)$$

Es decir, se trata de expandir las variables V_r a través de splines de regresión cúbicos restringidos, *RCS*, con q nudos $\{\xi_{r1}, \xi_{r2}, \dots, \xi_{rq}\}$.

Al no tener razones suficientes para suponer una ubicación concreta de los nudos, los fijamos de acuerdo la regla empírica que en general considera que se suelen situar entre 3 y 7 nudos. En caso de que el número de pacientes en la muestra sea inferior a 100, se colocarán menos de 5, mientras que si el número de pacientes es superior a 100, se suelen colocar 5 o más. Si el tamaño de la muestra es suficientemente grande ($N > 100$), se colocan el primero y el último nudos en los percentiles 5 y el 95 respectivamente.

Una tarea fundamental en la construcción del *SPANISH CARDIOVASCULAR RISK SCORECARD* consiste en encontrar las funciones de base que mejor especifican la no linealidad de las variables de riesgo cardiovascular $V_1, \dots, V_{p_2}$ que han manifestado no linealidad con respecto al *logit* de la probabilidad de *cardiohealth default.* El éxito de los modelos basados en la expansión lineal de funciones de base depende de la correcta especificación de las funciones base.

De acuerdo con la metodología planteada para la construcción del modelo, se partirá del modelo que notaremos por $M_{inicial}$, que se corresponde con el modelo logístico lineal en el que se incluyen todas las variables $U_1, \dots, U_{p_1}$ que han manifestado una relación de linealidad con el estado de *cardiohealth default* altamente significativo, con niveles de significación iguales o inferiores a $\alpha = 0{,}10$.

A continuación, siguiendo el segundo punto de la metodología constructiva propuesta, se procederá a introducir en el modelo las distintas expansiones de funciones de base

seleccionadas (constantes a trozos, pesos de la evidencia o splines cúbicos restringidos de Stone y Koo) para cada una de las variables $V_1, \dots, V_{p_2}$ de la componente no lineal del modelo.

Para cada una de las variables $V_j,\ j = 1, \dots, p_2$ se ajustarán los modelos alternativos por *LLR* y aquella expansión lineal por funciones de base de V_j para la que se consigan mejores estadísticos de bondad de ajuste, poder discriminante, poder predictivo y capacidad clasificatoria en combinación con la sencillez y la facilidad de interpretativa del modelo, será la que sustituya a V_j en el modelo logístico lineal híbrido.

- **Para la valoración de la bondad del ajuste del modelo a los datos de entrenamiento se calcularán los pseudo – coeficientes de determinación del modelo:**

Coeficiente de McFadden: $R^2(U)$ basado en la log – verosimilitud según la formulación siguiente:

$$R^2(U) = 1 - \frac{logver(M)}{logver(M_0)}$$

La aproximación de McFadden(65) pretende recoger para modelos lineales generalizados los conceptos del coeficiente de determinación R^2 como tasa de variabilidad explicada y como mejora del modelo nulo al modelo ajustado, respecto de este segundo aspecto, la razón de las verosimilitudes sugiere el nivel de mejora ofrecido por el modelo ajustado sobre el modelo con solo el intercepto. Por otro lado se verifica que $0 \leq R^2(U) < 1$.

Dado que una verosimilitud se sitúa entre 0 y 1, su logaritmo es menor o igual a cero. Si un modelo tiene una verosimilitud muy baja, entonces el logaritmo de la verosimilitud será mayor que el logaritmo de un modelo más verosímil. Por lo tanto, una proporción de logaritmo de verosimilitud muy baja indica que el modelo se ajusta mejor que el modelo con sólo intercepto. Si se comparan dos modelos sobre los mismos datos, el coeficiente $R^2(U)$ de McFadden será más alto para el modelo con la mayor verosimilitud.

El coeficiente $R^2(U)$ de McFadden adolece del problema que afecta a las log – verosimilitudes, aumentan con la introducción de complejidad en el modelo, por lo que

para evitar problemas de sobreajuste se puede adoptar el coeficiente $R^2(U)\_Ajustado$ de McFadden que penaliza la complejidad.

Coeficiente de Cox y Snell: Cox y Snell (1989) (66) presentaron una versión de $R^2(U)$ muy popular denotada por R^2, basada en la medida de verosimilitud $verosim(M)$. Por definición, $verosim(M)$ es la probabilidad de la variable dependiente dadas las variables independientes.

Cox y Snell definieron su pseudo coeficiente de determinación R^2, según la expresión:

$$R^2 = 1 - \left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}}$$

La relación de las verosimilitudes $\left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}}$, refleja la mejora del modelo completo sobre el modelo intercepto, de forma que cuanto sea menor esa cantidad, mayor será la mejora y tanto más se aproximará el $pseudo - R^2$ de Cox y Snell a 1. Por tanto, podemos decir que la aproximación de Cox y Snell se centra para modelos lineales generalizados en medir la mejora del modelo nulo al modelo ajustado.

El rango de este estadístico es el que se muestra en la siguiente expresión:

$$0 \leq R^2 \leq 1 - \left(verosim(M_0)\right)^{\frac{2}{N}}$$

El hecho de que el máximo del coeficiente R^2 de Cox y Snell pueda ser inferior a 1 conlleva que este estadístico sea difícil de interpretar.

Coeficiente de Nagelkerke: Nagelkerke (1991) (67) plantea una versión ajustada del coeficiente R^2 de Cox y Snell para asegurarse que valora entre 0 y 1. El $pseudo - R^2$ de Nagelkerke, que notaremos por $\widetilde{R}^2$, será normalmente mayor que el $pseudo - R^2$ de Cox y Snell, y adopta la forma:

$$\widetilde{R}^2 = \frac{1 - \left(\frac{verosim(M_0)}{verosim(M)}\right)^{\frac{2}{N}}}{1 - \left(verosim(M_0)\right)^{\frac{2}{N}}}$$

Si el modelo ajustado predice perfectamente los resultados y tiene verosimilitud 1, el $pseudo - R^2$ de Nagelkerke es $\widetilde{R}^2 = 1$. Sin embargo, si el modelo completo no mejora

al modelo con sólo intercepto, el $pseudo - R^2$ de Nagelkerke es mayor que cero, $\widetilde{R}^2 > 0$, por lo que el rango total $[0, 1]$ de los mínimos cuadrados ordinarios no está todavía cubierto.

Ninguno de los estadísticos $pseudo - R^2$ considerados aquí puede ser interpretado de forma independiente o comparando distintos conjuntos de datos, son válidos y útiles en la evaluación de varios modelos de predicción sobre la misma muestra de datos con la misma variable respuesta. En esta situación, un $pseudo - R^2$ más alto indica que el modelo predice mejor la respuesta.

Los test de bondad de ajuste que hemos visto hasta ahora basados exclusivamente en la log – verosimilitud negativa, conducen con frecuencia a rechazar modelos aceptables a la vez que a aceptar algunos que resultan menos parsimoniosos de lo que debieran, lo que va en contra de la capacidad de generalización y facilidad explicativa requerida al modelo.

Cuanto más complejo es el modelo mejor es el ajuste y por tanto, más alto es el valor de la verosimilitud que se obtiene. En otros términos esto significa que a mayor verosimilitud mayor es el sobreajuste, lo que hace necesarios otros criterios alternativos para medir la bondad del ajuste. Para resolver esta problemática se propusieron los criterios alternativos basados en la log – verosimilitud negativa penalizada, Criterio de Información de AKAIKE (AIC), y Criterio de Información Bayesiano de SCHWARZ (BIC), en los que el término de penalización es el encargado de corregir la complejidad del modelo.

El Criterio de Información de AKAIKE, para el modelo de regresión logística, usando la log – verosimilitud binomial, tiene la siguiente expresión:

$$AIC = -2logver + 2p$$

siendo p el número de variables explicativas de entrada.

AIC es asintótico por lo que se requieren muestras grandes. Además, el número máximo de parámetros no puede exceder $2pN$, donde N es el número de observaciones. Por último, resaltar que existen casos en los que AIC decrece monótonamente, es decir, no existe solución (en la mayoría de los casos el culpable es la mala selección del tipo de modelo).

Por otra parte, el Criterio de Información Bayesiana o Criterio de Schwarz al igual que AIC, es aplicable también en escenarios donde el ajuste se realiza a través de la maximización de la log – verosimilitud. La expresión genérica es:

$$BIC = -2logver + log(N)p$$

Para AIC y BIC cuanto menor es su valor, mejor es el ajuste. Dos de las mayores fortalezas de estas dos medidas son:

- Se puede comparar el ajuste de diferentes modelos incluso cuando los modelos no están anidados. Esto es particularmente útil cuando se tienen teorías que son muy diferentes. La idea básica es comparar la verosimilitud relativa de los dos modelos en vez de analizar la desviación absoluta de los datos observados de un modelo particular.
- Como medidas de información que son, penalizan la inclusión de variables que no mejoran significativamente el ajuste. En particular, con grandes muestras las medidas de información pueden conducir a modelos más parsimoniosos.

- **Para la valoración del poder discriminante del modelo** utilizaremos la medida de Área bajo la Curva ROC (AUC), donde tanto la curva como su estadístico asociado se obtienen sobre la muestra de validación. La valoración del poder discriminante es clave puesto que un modelo con alto poder discriminante es, sin duda alguna, un potente instrumento de predicción sobre la probabilidad asociada a la variable respuesta *cardiohealth default* en un horizonte temporal previamente fijado. En otras palabras, es un modelo con un alto porcentaje de aciertos frente a un bajo porcentaje de fallos.

La curva ROC permite visualizar el poder discriminante de un modelo de calificación, y es la que usualmente se utiliza en las aplicaciones prácticas de modelos de riesgo cuando el modelo proporciona la función de probabilidad, como es el caso del modelo *HLLM*. La Curva ROC consiste en la representación gráfica de los puntos de coordenadas $\{1 - F_0(c), 1 - F_1(c)\}$ para cada puntuación c, lo que formalmente podemos representar por la expresión:

$$ROC(u) = 1 - F_0[(1 - F_1)^{-1}(u)]\,,\ \ u \in [0,1]$$

siendo $F_0(c) = P(C \leq c|Y = 0)$ es la tasa de fallos para la puntuación c, llamada *especificidad*, y $F_1(c) = P(C \leq c|Y = 1)$ siendo $1 - F_1(c)$ la *sensibilidad*.

La medida del poder discriminante asociada a la Curva ROC, es el Área Bajo la Curva ROC, AUC.

El área bajo la curva ROC se obtiene a partir de la siguiente expresión:

$$AUC = \int_{+\infty}^{-\infty} \left[\left(1 - F_1(c)\right) d(1 - F_0)(c)\right]$$

Esta medida toma valores entre 0 y 1; 0 para la menor desviación y 1 para la mayor, si bien un AUC por debajo de 0,5 no tiene significado. Cuando el valor de AUC es de 0,5 significa que el modelo hace predicciones al azar. Un valor de AUC igual a 1 indica que las predicciones son perfectas. En general, cuanto mayor sea el área bajo la curva ROC mejor será el modelo.

El AUC es una transformación lineal de la Tasa de Precisión $AR = 2AUC - 1$, y puede interpretarse como la habilidad media del modelo de riesgo cardiovascular para clasificar exactamente a los pacientes según su estado de *cardiohealth default* en los que podrían presentar un evento cardiovascular y no presentarlo.

- **Para contrastar significación estadística de cada coeficiente β en el modelo**, utilizaremos el estadístico χ^2 de Wald, que prueba si la variable respuesta tiene una relación de dependencia significativa con cada variable explicativa, el estadístico de contraste viene dado por:

$$Z = \frac{\hat{\beta}}{SE}$$

siendo SE el error estándar. Según varios autores este estadístico para grandes coeficientes al aumentar el error estándar de manera significativa el valor $\chi^2\ de\ Wald$ se reduce.

Antes de considerar adecuada una expansión lineal por funciones de base de una variable para conseguir la significación en el modelo logístico lineal procedemos a la

evaluación del modelo resultante, de este modo podremos valorar si el modelo ajustado es un modelo válido y por tanto, lo es también la expansión lineal por funciones de base considerada, más allá de que presente un ajuste adecuado a los datos.

Una vez ajustados los datos, se obtienen los valores calculados de la función de calificación de pacientes y se estima el error de predicción sobre la muestra de validación. Será preferible, en lo que respecta a este apartado, el modelo con menor error de predicción sobre la muestra de validación

Los test basados en $-2logver$ conducen con frecuencia a rechazar modelos adecuados aceptando algunos que resultan menos parsimoniosos de lo que debieran. Por lo que será necesario utilizar aquí los criterios de Información de Akaike, AIC, y de Información Bayesiano de Schwarz, BIC, con el fin de que el término de penalización corrija la complejidad del modelo y se pueda evitar el sobreajuste.

Por otro lado, dado que las funciones de base se seleccionan con la intención de construir modelos de predicción y clasificación es evidente que AIC, BIC, AUC y el porcentaje de clasificación incorrecta se deben evaluar sobre la muestra de validación. Esta evaluación puede contribuir eficazmente a conocer la solidez de los indicadores calculados inicialmente sobre la muestra de entrenamiento y por tanto, pueden constituir una herramienta de inestimable ayuda en la selección de las funciones de base más prometedoras.

Como se ha comentado anteriormente, para cada una de las variables $V_j,\ j = 1, \dots, p_2$ se ajustan los modelos alternativos por LLR y aquella expansión lineal por funciones de base de V_j para la que se consigan mejores estadísticos de bondad de ajuste, de poder discriminante, poder predictivo y capacidad clasificatoria en combinación con la sencillez y la facilidad de interpretativa del modelo, será la que sustituya a V_j en el modelo logístico lineal híbrido. Este proceso se repite secuencialmente para cada $V_j,\ j = 1, \dots, p_2$ de la componente no lineal del modelo hasta obtener la estructura funcional del modelo final, M_{final}. A la vez se construirán modelos alternativos considerando expansiones lineales por funciones base que aunque no sean las más prometedoras en principio parezcan adecuadas, o al menos de interés a efectos comparativos.

Por último, en esta fase de especificación del modelo se valorará el ajuste desde el punto de vista estadístico del M_{final} y de los modelos alternativos. Esta valoración se hará a través de coeficientes de bondad de ajuste como el de Nagelkerke, el poder discriminante y explicativo mediante AUC, y la calidad como clasificadores mediante porcentajes de clasificados correctamente. La valoración anterior nos puede llevar a considerar más de uno de los modelos logísticos ajustados como correctos para explicar al estado de *cardiohealth default* desde el punto de vista estadístico.

Con la metodología expuesta el modelo ***LPLM*** (6.21) se derrumba dando lugar al modelo ***HLLM*** constructivo anterior (o a los modelos alternativos), herramienta de predicción y clasificación capaz de explicar la relación de dependencia de la variable estado de *cardiohealth default* con las variables de riesgo cardiovascular más relevantes, estimar la función de probabilidad de *default*, calificar a los pacientes mediante su puntuación cardiovascular, explicar la puntuación que se les otorga y clasificar a un nuevo paciente con la precisión adecuada.

7.5. EVALUACIÓN, GENERALIZACIÓN Y SELECCIÓN DEL MODELO

Antes de proceder a la selección del modelo final y fijar su error de predicción o generalización esperado, es necesario evaluar los distintos modelos alternativos con el fin de detectar cuál de ellos reúne las cualidades idóneas.

En este sentido, además de estimar el error empírico del modelo sobre la muestra de validación, error de evaluación, se han de evaluar posibles errores de especificación de la componente sistemática, de la distribución de probabilidad de la componente aleatoria y de la relación asumida entre ambas componentes del modelo en la fase de especificación.

Igualmente es necesario valorar las posibles pérdidas de eficacia estadística al priorizar la facilidad explicativa de las transformaciones de tramado de variables ya sea por funciones constantes a trozos obtenidas a partir de indicadores de particiones recursivas, por pesos de la evidencia asignados a tramados de las variables o por splines cúbicos restringidos.

Como resultado de este proceso se obtendrá una aproximación al modelo más prometedor dado nuestro objetivo, que será el considerado como ***modelo óptimo***. A partir de este modelo óptimo se podrá tanto estimar la probabilidad de *cardiohealth default*, como la función de calificación de pacientes y el clasificador de nuevos pacientes susceptibles de recibir una valoración de su riesgo cardiovascular.

- **Evaluación de los modelos HLLM preseleccionados**

La evaluación del modelo supone valorar si el modelo ajustado en la etapa de estimación es un modelo válido, más allá de que presente un ajuste adecuado a los datos. Se propone evaluar los modelos a través del *Criterio de Información Bayesiano*, BIC, puesto que seleccionar el modelo con mínimo BIC es equivalente a elegir el modelo con mayor probabilidad a posteriori, y analizar el poder discriminante de los modelos a través el *Área bajo la curva ROC*, AUC.

- **Generalización. Error Test de los modelos HLLM preseleccionados.**

Con el objetivo de conocer si los modelos alternativos seleccionados son suficientemente generalizables, así como establecer su ranking respecto de esta característica, se debe analizar el error de predicción esperado sobre la muestra test para cada uno de ellos, es decir el error test. Para ello se utilizará la muestra test, que es independiente de las muestras de entrenamiento y validación.

El error test empírico no es un buen estimador del error de generalización. Sin embargo, si el conjunto test ha sido elegido aleatoriamente, la ejecución del modelo sobre el conjunto de entrenamiento y el error sobre el conjunto test proporcionará un estimador insesgado del error de generalización.

A partir de las medidas propuestas para evaluar el modelo, valorar su capacidad de generalización y bajo el requerimiento de facilidad de interpretación, se seleccionará el modelo más adecuado para llevar a cabo la estimación del riesgo cardiovascular.

- **Selección del modelo de riesgo cardiovascular**

Se elaborará un ***Ranking del Ajuste de los Modelos*** para la selección del modelo óptimo. Este ranking consistirá en la comparativa entre los modelos propuestos de la bondad de los modelos a través del Criterio de Información Bayesiano BIC, del poder

discriminante mediante el Área bajo la Curva ROC, AUC ambos sobre la muestra de validación y del Error Esperado de Generalización sobre la muestra.

Además de estas medidas, si fuera necesario se podrían incluir otras como el pseudo – coeficiente de Nagelkerke o el Criterio de Información de AKAIKE (AIC).

La selección del mejor modelo de estimación para el pronóstico de la relación de dependencia entre el estado de *cardiohealth default* y las variables explicativas del modelo consideradas, se llevará a cabo mediante un equilibrio entre el *Ranking del Ajuste del Modelo* y la valoración del requerimiento relativo a la facilidad de interpretación, principalmente para la parte no lineal que es la que ha sido expansionada por funciones de base.

- **Sensibilidad de la selección de las funciones de base a la configuración de la muestra de entrenamiento**

Una vez se han obtenido las funciones de base más prometedoras para la estimación del modelo logístico y antes de iniciar la búsqueda del modelo más parsimonioso, se debe comprobar si el método seguido para su obtención es o no excesivamente sensible a la configuración de la muestra de entrenamiento.

Una forma de abordar esta cuestión consiste en analizar la estabilidad de los parámetros y estadísticos de bondad de ajuste, poder discriminante y clasificación correcta frente al muestreo de entrenamiento, lo que se puede realizar a través del ajuste por regresión logística *bagged* del modelo seleccionado sobre n $(n \geq 200)$ *remuestras bootstraps* de la muestra de entrenamiento.

La propuesta para la sensibilidad del método puede basarse en las discrepancias entre los estimadores de los coeficientes del modelo y los diferentes indicadores de la bondad del ajuste, del poder discriminante y de la eficacia de clasificación del mismo ajustado sobre la muestra de entrenamiento, ***HLLR***, y los promedios de los n ajustes sobre la *remuestras bootstraps*, ***HLLR_Bag***

- **Reducción de la complejidad del modelo**

Para reducir la complejidad del modelo y evitar posibles problemas como el sobreajuste, se procederá a aplicar técnicas de regularización o *poda* para reducir el número de funciones de base.

Se propone el *Proceso de Selección Limitada de Funciones de Base hacia Atrás* hasta conseguir un ***modelo óptimo reducido***. Se ha de verificar que este modelo reducido presenta buenas condiciones estadísticas; es decir que sea un modelo caracterizado por un buen ajuste (coeficiente de Nagelkerke), un alto poder discriminante y predictivo (Área bajo la curva ROC), una alta eficacia como clasificador (% de clasificados correctamente) y que generaliza bien.

7.6. PODER DISCRIMINANTE

La puntuación $C(X)$ asignada a los pacientes sintetiza la información proporcionada por un conjunto de variables seleccionadas por su influencia en el comportamiento de la variable *cardiohealth default*, que consideramos indicadora de la salud cardiovascular del paciente. Por ello a cada individuo se le asocian dos variables, $C(X)$ que representa una puntuación sobre una escala continua que le asigna el sistema de puntuación al riesgo del paciente y la variable Y, que muestra el estado de *cardiohealth default* o *no default* que el paciente presenta al final del horizonte temporal fijado que será de un año.

El objetivo que se persigue a través de $C(X)$ es el de pronosticar el futuro estado de Y para el paciente, confiando en la información sobre el riesgo cardiovascular contenida en $C(X)$. La función de calificación cardiovascular asignará puntuaciones altas a aquellos pacientes que presenten alta probabilidad de presentar un evento cardiovascular, y puntuaciones bajas a aquellos pacientes que tengan una baja probabilidad de presentar un evento cardiovascular.

Una característica básica que han de presentar estas funciones es su alta eficiencia para separar a los pacientes en dos grupos atendiendo a su riesgo, por tanto, las medidas del poder discriminante pueden usarse como medidas de la eficacia del sistema de puntuación.

El objetivo de esta fase es el de juzgar si la función de calificación o función de puntuación cardiovascular asociada al modelo es apropiada para discriminar entre los pacientes que presentan evento cardiovascular y los que no lo presentan, para ello se califica a través del modelo *HLLM* a la totalidad de los individuos de la muestra y se valida si los estadísticos de poder discriminante son suficientemente significativos.

Se analizará el *Perfil de la diferencia entre las funciones de distribución acumulativas* de las poblaciones de buenos y malos, y su estadístico de Kolmogorov – Smirnov asociado. La conveniencia del uso de estos dos instrumentos estadísticos radica por un lado en que el perfil de las funciones de distribuciones de las poblaciones de *default* y *no default* es un instrumento gráfico muy potente para visualizar las posibles diferencia entre ambas poblaciones, y por otro, en que se cuenta con un test asociado para contrastar si las diferencias son o no significativas.

A continuación, se estudian otras medidas del poder discriminante tales como la *Tasa de Precisión* AR, que se obtiene a partir del *Índice de Gini* resumen de la *Curva de Ajuste Acumulativo* CAP o *Curva de Lorenz*.

7.6.1. Perfil de la diferencia entre las funciones de distribución acumulativas. Test estadístico de Kolmogorov-Smirnov asociado

Se analiza el *Perfil de la diferencia entre las Funciones de Densidad y Distribución acumuladas* para las poblaciones de los que presentan y no presentan el evento valorado, y su estadístico de Kolmogorov – Smirnov asociado. La conveniencia del uso de estos dos instrumentos estadísticos radica por un lado, en que el perfil de las funciones de distribuciones de las poblaciones de *cardiohealth default* y *no cardiohealth default* es un instrumento gráfico muy potente para visualizar las posibles diferencias entre ambas poblaciones, y además se cuenta con un test asociado para contrastar si las diferencias son o no significativas.

Denotaremos por $F_0(\cdot)$ la función de distribución acumulada de $C(X)|Y=0$, es decir de la función de calificación condicionada al grupo de los pacientes que no presentan el evento valorado, y análogamente $F_1(\cdot)$ la función de distribución acumulada de $C(X)|Y=1$, es decir de los pacientes que presentan el evento. Sean $f_0(\cdot)$ y $f_1(\cdot)$ las correspondientes funciones de densidad, es decir:

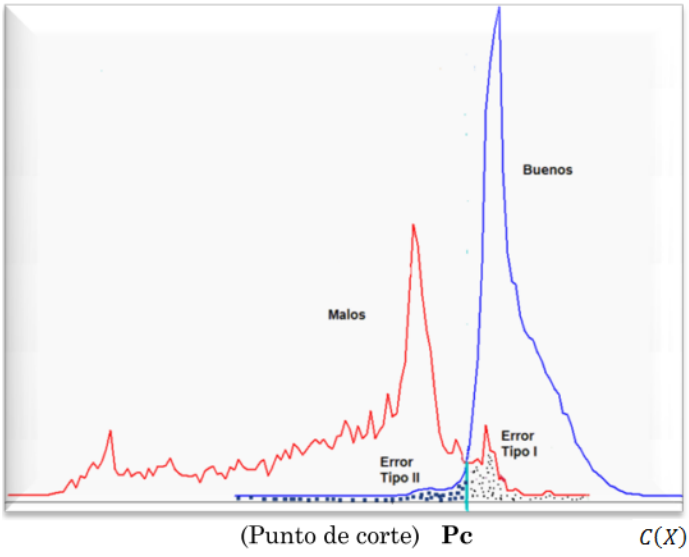
$$F_0(c) = P(C \leq c|Y = 0) = \int_{-\infty}^{c} f_0(u)du$$

$$F_1(c) = P(C \leq c|Y = 1) = \int_{-\infty}^{c} f_1(u)du$$

$$F(c) = P(C \leq c|Y = 0) = \int_{-\infty}^{c} f(u)du$$

Para el modelo ***HLLM*** se representan gráficamente las funciones de densidad de las puntuaciones asignadas por la función de calificación de pacientes de ambas poblaciones (presentan y no presentan el evento) estimadas a través de la estimación no paramétrica por el núcleo Gaussiano.

A continuación se fija un punto de corte p_c, frontera de separación entre los niveles de calificación de los pacientes, de forma que aquellos individuos cuya puntuación esté por debajo de ese valor son calificados como "candidatos a presentar un evento cardiovascular" y los que tengan un valor superior se califiquen como "no candidatos a presentar un evento cardiovascular". La elección de este punto de corte que puede ser aquel tal que $f_0(p_c) = f_1(p_c)$, es decir la abscisa del punto de corte de las funciones de densidad de ambas poblaciones, frontera entre dos áreas de solapamiento de ambas curvas. Como alternativa a la elección del punto de corte anterior se podría plantear aquella basada en algún criterio de probabilidad.



Una vez determinado el punto p_c, se elabora la *Matriz de Confusión* recogiendo el número de individuos de la muestra según su estado de *cardiohealth default* estimado y observado.

		ESTIMADOS	
		Buenos	***Malos***
OBSERVADOS	***Buenos***	a	b
		$Verdadedo$	$Falsa\ Alarma$
	Malos	c	d
		$Fallo$	$Éxito$

Existen dos casos para los que dado el punto de corte p_c, ocurre una buena predicción del comportamiento de un paciente en cuanto al cumplimiento o no de la salud cardiovascular. Aclarar que se entenderá un paciente presenta cumplimiento de la salud cardiovascular si no aparece el evento en el horizonte de un año prefijado y no cumplimiento en caso contrario. Los casos de buena predicción son:

1. Al principio del horizonte temporal de un año, el modelo predijo para un paciente incumplimiento, y durante dicho período el paciente incumple. En este caso se dice que se ha tenido un *éxito*.
2. Cuando el modelo predice cumplimiento antes de comenzar el período y el paciente no incumple a lo largo del período.

También se observan dos casos de predicción errónea:

1. Una primera situación de mala predicción se presenta si para un punto de corte, frente a la predicción de incumplimiento por parte del modelo, el paciente no presenta actualmente ningún incumplimiento (Tasa de Falsa Alarma, Tasa β o Error Tipo II) y que se obtiene en la forma siguiente:

$$Error\ tipo\ II = \frac{b}{b+d}$$

   El Error tipo II muestra el porcentaje de pacientes de la muestra considerados en riesgo cuando en realidad pertenecen a la población que no está en riesgo de presentar el evento cardiovascular valorado.

2. Una segunda situación de mala predicción se presenta cuando se estimó cumplimiento y en la actualidad el paciente incumple la salud cardiovascular por

haber presentado el evento en el intervalo de tiempo considerado. Es el llamado Fallo o Error de Tipo I, que se obtiene como sigue:

$$Error\ tipo\ I = \frac{c}{a + c}$$

El Error tipo I facilita el porcentaje de pacientes que el modelo consideraría que no presenta riesgo de presentar un evento cuando en realidad es un paciente de la población de riesgo.

La **proporción de éxitos** o **tasa de aciertos** *HR* (*Hit Ratio*), es el porcentaje de pacientes que estimó el modelo como *cardiohealth default* al principio del horizonte temporal y que con el punto de corte considerado p_c presentaron el evento cardiovascular y por tanto la estimación fue correcta.

La **proporción de falsas alarmas** *FAR (False Alarm Ratio)*, es el porcentaje de pacientes estimados por el modelo como *cardiohealth default* para el punto de corte p_c y que resultaron *no cardiohealth default* al final del horizonte temporal.

El **error total de clasificación** es la suma del error de tipo I y el error tipo II.

El **error total de clasificación** es uno de los criterios propuestos para medir el rendimiento del modelo en la clasificación de pacientes en riesgo.

La prueba del poder discriminante se debe valorar en un conjunto de datos independientes fuera de la muestra de validación. De lo contrario existe el peligro de que el poder discriminante pueda ser exagerado por el sobreajuste en el conjunto de datos de entrenamiento. Si el modelo presenta un poder discriminante relativamente bajo en un conjunto de datos independiente aunque estructuralmente similar al del conjunto de datos de entrenamiento, significa que el modelo tiene una baja estabilidad. Un rasgo característico de un modelo estable es que recoge de forma adecuada la relación causal entre los factores de riesgo y el *cardiohealth default*, lo que evita dependencias espurias derivadas de correlaciones empíricas. En contraste con los modelos estables, los sistemas inestables con frecuencia muestran una disminución considerable en el nivel de precisión de la estimación con el tiempo.

La propuesta de medición del poder discriminante de un modelo se ha elegido por la simplicidad que presentan sus cálculos, pero en general no es un criterio apropiado para medir el rendimiento de un modelo de clasificación. Por tanto, será necesario utilizar

otras medidas alternativas para medir el poder discriminante del modelo conjuntamente con los test estadísticos correspondientes. Una medida importante se basa en la región de solapamiento de las funciones de densidad que se describe a continuación.

La región de solapamiento O está constituida por la zona bajo la densidad de *buenos*, a la izquierda del umbral p_c, y la zona bajo la densidad de *malos* a la derecha de p_c. Dado que hay un solo punto de intersección óptimo, y la relación entre la puntuación $C(X)$ y la probabilidad de incumplimiento es monótona positiva, el área de solapamiento está definida por:

$$O = min_{p_c}\{F_1(p_c) + 1 - F_0(p_c)\} \quad (7.14)$$

Por lo que se puede definir una medida del poder discriminante como:

$$\left|D_{N_0N_1}\right| = 1 - O = max|F_0(p_c) - F_1(p_c)| \quad (7.15)$$

El indicador del poder discriminante $\left|D_{N_0N_1}\right|$ toma valores en el intervalo $[0,1]$, donde $\left|D_{N_0N_1}\right| = 1$ indica una separación total y $\left|D_{N_0N_1}\right| = 0$ significa que no existe separación alguna.

Si se considera la hipótesis donde la hipótesis nula $F_1(p_c) = F_0(p_c)$, que indica que las dos muestras resultan de la misma población, es decir, la función de calificación de pacientes no discrimina entre *cardiohealth default* y no *cardiohealth default*, o de forma más general, la variable de calificación $C(X)$ no influye sobre la probabilidad de *default*, entonces se rechaza la hipótesis nula a nivel de significación α si:

$$T = \left|D_{N_0N_1}\right| \sqrt{\frac{N_0N_1}{N_0+N_1}} > \kappa_{N_0,N_1,1-\alpha/2} \quad (7.16)$$

donde κ_α se obtiene de $P(\kappa \leq x) = 1 - \alpha$, donde κ es una variable aleatoria con función de distribución acumulada:

$$P(\kappa \leq x) = 1 - \sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}e^{-\frac{(2i-1)^2\pi^2}{8x^2}} \quad (7.17)$$

7.6.2. Curva de Ajuste Acumulativo, *CAP*. Tasa de Precisión, *AR*

La *Curva de Ajuste Acumulativo CAP o Curva de Lorenz* representa gráficamente la distribución de la función de calificación frente a las distribuciones condicionadas de

cardiohealth default, lo que posibilita la comparación gráfica de diferentes puntuaciones de riesgo. Se representan gráficamente los puntos de coordenadas $\{1 - F(c),\ 1 - F_1(c)\}$ para cada puntuación. La cantidad $F(c) = P(C \leq c)$ coincide con la Tasa de Alarma para la puntuación c y $F_1(c) = P(C \leq c|Y = 1)$ coincide con la Tasa de Aciertos para la misma puntuación. La cantidad $100 \times L(c)$ indica el porcentaje de pacientes en riesgo que se hallan entre los $100 \times c$ primeros pacientes de acuerdo con sus puntuaciones.

Una *Curva de Lorenz* idéntica a la diagonal corresponde a una puntuación que ordena a los pacientes de forma totalmente aleatoria, obteniéndose por tanto un modelo sin ningún poder discriminante. El modelo de calificación real está entre los dos extremos, y por tanto, el área de la región comprendida entre la CAP y el modelo aleatorio es una buena medida de la eficacia de la puntuación. Uno de esas medidas es el *Coeficiente de Gini* G, que consiste en dos veces esa área.

La medida más usual y popular sobre la precisión de un modelo de riesgo, alternativa al *Coeficiente de Gini*, es la *Tasa de Precisión* AR también basada en la CAP.

La *Tasa de Precisión* AR es el cociente entre el *Índice de Gini* de la curva CAP y el *Coeficiente de Gini* de la *Curva de Lorenz Optimal* que se corresponde con la puntuación que separa perfectamente a los pacientes que no están en riesgo de los que están en riesgo de presentar el evento cardiovascular. Esta curva tiene la peculiaridad de que la ordenada $(1 - F_1)(c) = 1$ se alcanza para la abscisa $c = P(Y = 1)$, por lo que para la *Curva de Lorenz Optimal* se tiene que el *Coeficiente de Gini Optimal*:

$$G_{opt} = P(Y = 0) = 1 - P(Y = 1)$$

La *Tasa de Precisión* AR, viene dada por la relación del *Coeficiente de Gini* de cada puntuación y el *Coeficiente de Gini Optimal*.

$$AR = \frac{G}{G_{opt}}$$

El valor de AR se sitúa entre 0 y 1 si la *Curva de Lorenz* es realmente cóncava, es decir, si existe una relación monótona positiva entre $C(X)$ e Y. El modelo de calificación óptimo es aquel para el que más se aproxime AR a 1. El peor modelo de calificación es aquel para el que más se aproxime AR a 0, es decir, el que más se aproxime al modelo

aleatorio. Esto hace que AR sea no sólo una medida idónea para medir el poder discriminante de un modelo, sino también para comparar diferentes puntuaciones.

7.6.3. Curva *ROC*. Área bajo la Curva *ROC, AUC*

La curva que permite visualizar el poder discriminante de un modelo de calificación es la curva ROC siendo la medida del poder discriminante asociada el área bajo la curva ROC, AUC.

Otra curva que permite visualizar el poder discriminante de un modelo de calificación es la curva ROC. Esta es la que usualmente se utiliza en las aplicaciones prácticas de modelos de riesgo en los casos en que éste proporciona la función de probabilidad, como es el caso del modelo ***HLLM***. Esta curva de aspecto muy similar a la curva CAP presenta frente a ésta una importante diferencia, ya que en el eje horizontal se sitúa $1 - F_0(c)$ mientras en CAP se sitúa $1 - F(c)$, siendo $F(c)$ la función de distribución acumulada del total de pacientes.

La Curva ROC consiste en la representación gráfica de los puntos de coordenadas $\{1 - F_0(c),\ 1 - F_1(c)\}$ para cada puntuación c.

La medida del poder discriminante asociada a la *Curva* ROC, es el área bajo la curva AUC, cuanto mayor sea el área bajo la *Curva* ROC mejor será el modelo.

7.6.4. Test U de Mann – Witney

Dado que no existe un valor mínimo para AUC con significación estadística que nos permita decidir si el modelo de calificación tiene bastante poder discriminante, se plantea utilizar el test U de Mann – Whitney para rechazar significativamente la hipótesis nula de que el modelo no tiene más poder discriminante que el modelo aleatorio.

El test U para funciones de calificación de pacientes en su forma más simple se puede deducir en la forma siguiente:

Si denotamos por c_{j0} todas las puntuaciones observadas de *no cardiohealth default* y por c_{i1} todas las puntuaciones observadas de *cardiohealth default*, el estadístico del test U viene dado por:

$$\widehat{U} = \#\{c_{i1} > c_{j0}\} \;\; sobre\ todo\ i,j. \qquad (7.18)$$

Para una separación perfecta de pacientes *cardiohealth default y no default,* se obtiene $\widehat{U} = N_0 N_1$.

Si C e Y no están totalmente relacionadas, entonces el suceso $c_{i1} > c_{j0}$ ocurre con probabilidad $1/2$ de forma que $U \approx N_0 N_1$. En consecuencia, una versión escalada del estadístico $\widehat{U}$ será $\widetilde{U} = \frac{\widehat{U}}{N_0 N_1}$, es un estimador para el área bajo la curva que se puede obtener según la siguiente expresión:

$$U = P[(C|Y=1) > (C|Y=0)] = \int_{+\infty}^{-\infty} \left[\left(1 - F_1(c)\right) d(1 - F_0)(c)\right] = AUC \qquad (7.19)$$

Y por lo tanto, dado que $AR = 2AUC - 1$, se tiene que:

$$U = \left(\frac{AR+1}{2}\right) N_0 N_1 \qquad (7.20)$$

Se demuestra que bajo la hipótesis $F_1(c) = F_0(c)$, con N_0 y N_1 grandes, U se distribuye aproximadamente normal con media $\mu_U = \frac{N_0 N_1}{2}$ y desviación típica $\sigma_U = \sqrt{\frac{N_0 N_1 (N_0 + N_1 + 1)}{12}}$, por tanto:

$$Z = \frac{U - \frac{N_0 N_1}{2}}{\sqrt{\frac{N_0 N_1 (N_0 + N_1 + 1)}{12}}} \sim N(0,1) \qquad (7.21)$$

El test U de Wilcoxon – Mann – Witney se construye entonces a partir de:

Test	H_0	H_1	***Test Estadístico***	***Rechazo***
(1)	$F_1(c) = F_0(c)$	$F_1(c) > F_0(c)$	U	$U > k_{N_1,N_0,1-\alpha}$
(2)	$F_1(c) = F_0(c)$	$F_1(c) < F_0(c)$	U	$U < N_1 N_0 - k_{N_1,N_0,1-\alpha}$

Siendo el valor crítico,

$$k_{N_1,N_0,1-\alpha} = \frac{N_0 N_1}{2} + Z_{1-\alpha}\sqrt{\frac{1}{12}N_0 N_1 (N_0 + N_1 + 1)} \qquad (7.22)$$

Por último, dado que las propiedades estadísticas de AUC coinciden con las del estadístico de Mann – Witney, podemos aplicar el potente Test U de Wilcoxon – Mann – Witney para comparar el AUC del modelo ***HLLM*** con el del modelo aleatorio, $AUC = 0{,}5$. Dos de los test no paramétricos más clásicos para contrastar si dos distribuciones son o no idénticas son el test de la suma de rangos de Wilcoxon y su equivalente, el test U de Mann – Witney.

7.7. CALIBRACIÓN DEL MODELO

En la práctica las probabilidades de *cardiohealth default* pronosticadas difieren de las tasas de riesgo observadas finalmente. El problema surge cuando estas desviaciones no se producen al azar, sino sistemáticamente, lo que nos llevaría a pensar que el modelo de riesgo no sea el adecuado.

La cuestión que abordaremos en este apartado es "cómo la probabilidad de *cardiohealth default* pronosticada por el modelo de riesgo al comienzo del horizonte temporal puede ser revisada dadas las tasas de *cardiohealth default* realmente observadas al final de dicho período", a esta revisión se la denomina *"calibración del modelo"*.

La fase de calibración del modelo es clave en la propuesta del sistema de cuantificación *Spanish Cardiovascular Risk Scorecard* y marca una diferencia metodológica fundamental con respecto a las tablas de estimación del riesgo cardiovascular actuales.

Las tablas de riesgo cardiovascular de las que se dispone en la actualidad, estiman la probabilidad de presentar el evento cardiovascular valorado con un horizonte temporal muy amplio, que habitualmente es de diez años. Este planteamiento conlleva una serie de limitaciones claras. Una de ellas es la falta de precisión en la estimación a largo plazo, debido a que la estimación de lo que ocurrirá durante los diez años siguientes se realiza a partir de los valores actuales de las variables explicativas del riesgo, sin ningún

tipo de corrección por la modificación que dichos valores puedan sufrir a lo largo del amplio periodo de tiempo considerado. Otra de las limitaciones del planteamiento actual de las tablas de estimación sobre la que consideramos es importante hacer referencia, es la falta de especificación de la enorme diferencia que existe entre que el episodio ocurra en el mes siguiente o lo haga a los nueve años y once meses. A partir de las tablas no es posible saber el momento en el cual se producirá el evento valorado, sólo es posible saber que ocurrirá con una determinada probabilidad a lo largo de los diez años siguientes.

El planteamiento que se hace en esta Tesis marca un punto de inflexión con respecto a la perspectiva de estimación de las tablas de riesgo actuales. La propuesta de estimación de la probabilidad de *cardiohealth default* se realiza con un horizonte temporal de un año, momento en el cual se estudiará la necesidad de calibración del modelo para así asegurar la precisión de las estimaciones obtenidas a partir del mismo.

La calibración dota de capacidad al modelo para realizar estimaciones no sesgadas (objetivas) de las probabilidades de *cardiohealth default*. Por ello, decimos que un modelo está bien calibrado cuando la probabilidad de *cardiohealth default* pronosticada por el modelo de riesgo ***HLLM*** se desvía solo marginalmente de las tasas de *cardiohealth default* que han sido observadas. Por tanto, la calibración compara las probabilidades de *cardiohealth default* pronosticadas al comienzo del horizonte temporal de referencia con las tasas de *cardiohealth default* observadas al final del período, analizando las discrepancias entre unas y otras en orden a discernir si las mismas se deben a factores sistemáticos o aleatorios. En definitiva, se plantea el análisis de si los hechos acaecidos a posteriori respaldan los pronósticos a priori, de no ser así seguramente el modelo no sea el más adecuado.

El grado de discrepancia entre la probabilidad de *cardiohealth default* pronosticada y las tasas de *cardiohealth default* realmente observadas puede indicar problemas potenciales y acciones que necesitan ser acometidas.

Para la calibración, se asignarán en primer lugar categorías de calificación de los pacientes. Estas categorías de calificación se obtendrán a través de alguna regla con respecto a la probabilidad de *cardiohealth default*. Por ejemplo, mediante la probabilidad de *cardiohealth default* media por categoría, o por división en intervalos de determinado tamaño, o también utilizando el algoritmo *CART* para generar las

categorías, todo ello de modo que los individuos estén razonablemente distribuidos a través de estas categorías, sin concentraciones excesivas. Este aspecto de asignación de categorías de calificación es un campo abierto de investigación que deberá ser supervisado tanto por los estadísticos como por los expertos en riesgo cardiovascular.

Se plantean estadísticos de tipo *backtesting* para contrastar la siguiente hipótesis:

H_0: *La probabilidad de cardiohealth default pronosticada en un nivel de calificación es correcta.*

H_1: *La probabilidad de cardiohealth default pronosticada en un nivel de calificación es incorrecta.*

Para la formalización del test consideramos un sistema de riesgo con N pacientes clasificados en R categorías de calificación diferentes de acuerdo con sus calificaciones de riesgo cardiovascular. Si N_r indica el número de pacientes que son clasificados en la clase de clasificación r siendo $r \in \{1, \dots, R\}$, se tiene que $N = \sum_{r=1}^{R} N_r$ e indicando por:

$0 < \hat{p}_r < 1$: Probabilidad de *cardiohealth default* pronosticada por el sistema de calificación para la categoría r.

$0 < p_r < 1$: Probabilidad de *cardiohealth default* real (desconocida) para la categoría r.

$0 < p_r^{obs} < 1$: Tasa de *cardiohealth default* observada para la categoría r.

El test de hipótesis se puede plantear mediante dos formulaciones, test de una cara o test de dos caras:

Test de una cara	***Test de dos caras***
$H_0: P_1 = \hat{P}_1, \dots, P_R = \hat{P}_R$	$H_0: P_1 = \hat{P}_1, \dots, P_R = \hat{P}_R$
$H_1: \exists r \in \{1, \dots, R\}\ con\ P_r > \hat{P}_r$	$H_1: \exists r \in \{1, \dots, R\}\ con\ P_r \neq \hat{P}_r$

El test estadístico puede usarse para un cierto nivel de significación prefijado con su correspondiente $p - valor$, para tomar una decisión con ese nivel de significación. Altos $p - valores$ indican que el test es significativo y, por tanto, no se rechaza la hipótesis nula de que la probabilidad de *cardiohealth default* es al nivel p significativamente infra estimada. La elección del apropiado nivel de significación depende del grado de precisión y conservadurismo que se pretenda en cada caso.

Los métodos propuestos para la calibración de la probabilidad de *cardiohealth default* sobre un único periodo de tiempo bajo la hipótesis de independencia de los sucesos de *default* son el ***Test Binomial*** (Engelmann y Rauhmeir, 2006), el ***Test de Hosmer – Lemeshow*** (χ^2) (Hosmer y Lemeshow, 2000), y el ***Test de Spiegelhalter*** (Spiegelhalter, 1986). El *Test Binomial* sólo se puede aplicar a un simple grado de calificación sobre un único periodo de tiempo, mientras que los otros dos proporcionan métodos más avanzados que pueden usarse para contrastar la adecuación de la predicción a la probabilidad de *cardiohealth default* sobre un único periodo de tiempo para varias categorías de calificación de pacientes.

7.7.1. Test Binomial

El *Test Binomial* está diseñado para contrastar los pronósticos de la probabilidad de default estimada por el modelo $\hat{P}$, frente a la tasa de default observada P^{obs}, para una categoría de calificación r dada usando el siguiente test:

H_0: $P_r = \hat{P}_r$	La probabilidad de *cardiohealth default* real coincide con la estimada en la categoría r.
H_1: $P_r > \hat{P}_r$	La probabilidad de *cardiohealth default* es infra estimada en la categoría r.

Si se asume que los default ocurren independientemente para cada categoría r, se busca contrastar si la probabilidad de default de una categoría de calificación es correcta frente a la alternativa de que está infra estimada, es decir contrastar la hipótesis nula frente a la alternativa de una cara. Para un nivel de significación α se rechaza la hipótesis nula si el número de *defaults* $N_r P_r^{obs}$, es mayor que un valor crítico dado por:

$$k^* = min\left\{ \frac{k}{\sum_{i=k}^{N_r} \binom{N_r}{i} \hat{p}_r (1-\hat{p}_r)} \leq \alpha \right\} \qquad (7.23)$$

Siendo N_r el número de pacientes de la categoría r. Dado que para grandes valores de N_r el cálculo de k^* según (7.23) es costoso, se suele usar el hecho de que la distribución Binomial converge a la distribución *Normal* cuando el número de pruebas crece,

$$P_r^{obs} \sim N\left(\hat{P}_r, \frac{\hat{P}_r(1-\hat{P}_r)}{N_r}\right) \qquad (7.24)$$

lo que equivale a,

$$z = \frac{P_r^{obs} - \hat{P}_r}{\sqrt{\frac{\hat{P}_r(1-\hat{P}_r)}{N_r}}} \sim N(0,1) \qquad (7.25)$$

Se rechaza la hipótesis nula si la tasa de default observada $\hat{p}_r^{obs}$ es mayor que $p_{1-\alpha}$, siendo:

$$p_{1-\alpha} \approx \Phi^{-1}(1-\alpha)\sqrt{\frac{\hat{p}_r(1-\hat{p}_r)}{N_r}} + \hat{p}_r \qquad (7.26)$$

Esta aproximación de la distribución Binomial a la distribución *Normal* se aplicará cuando $N_r > 1000$.

Para el test de dos caras:

H_0: $P_r = \hat{P}_r$ La probabilidad de *cardiohealth default* real coincide con la estimada en la categoría r.

H_1: $P_r \neq \hat{P}_r$

El estadístico de contraste será también $N_r P_r^{obs}$, y se tiene que la región crítica para p_r^{obs} y un nivel de significación asintótica α está dado por:

$$[0, p_{\alpha/2}) \cup (1 - p_{\alpha/2}, 1] \qquad (7.27)$$

Por tanto, se rechazará la hipótesis nula para la categoría de clasificación r, si la tasa de *cardiohealth default* observada p_r^{obs} queda fuera del intervalo (7.27) calculado.

7.7.2. Test de Hosmer – Lemeshow

El test binomial (o su extensión normal) es adecuado para contrastar un único nivel de calificación, pero no de varios o todas las categorías de calificación simultáneamente. El Test de Hosmer – Lemeshow o Test Chi – cuadrado, es en esencia un test de conjunto para varios grados de calificación.

Se busca contrastar si las probabilidades de *cardiohealth default* son correctas para todas las categorías de calificación de pacientes simultáneamente, es decir, contratar:

$$H_0: P_1 = \hat{P}_1, \dots, P_R = \hat{P}_R$$

$$H_1: \exists r \in \{1, \dots, R\} \; con \; P_r \neq \hat{P}_r$$

Se asumen las siguientes hipótesis:

i. Las probabilidades de cardiohealth default pronosticadas por el modelo $\hat{p}_r$, y las tasas de default observadas p_r^{obs} son idénticamente distribuidas.

ii. Todos los sucesos de cardiohealth default tanto dentro de cada categoría como entre las categorías son independientes.

El test estadístico chi – cuadrado se deduce del estadístico chi – cuadrado de Pearson original y viene dado por:

$$t_R = \sum_{r=1}^{R} N_r \frac{(p_r^{obs} - \hat{p}_r)^2}{\hat{p}_r(1-\hat{p}_r)} \qquad (7.28)$$

Bajo las hipótesis *i* e *ii.*, cuando $N_r \to \infty$ simultáneamente para todo $r = 1, \dots, R$, por el Teorema Central del Límite se tiene que la distribución de t_R converge en distribución a una distribución χ^2 con R grados de libertad:

$$t_R = \sum_{r=1}^{R} N_r \frac{(P_r^{obs} - P_r)^2}{\sqrt{P_r(1-P_r)}} \xrightarrow{D} \chi^2(R) \qquad (7.29)$$

Por tanto, se rechazará la hipótesis nula para un nivel de significación asintótico α, si t_R es mayor que el $(1-\alpha) - cuantil$ de la distribución χ^2 con R grados de libertad.

El $p - valor$ del test χ^2 es una medida para validar la adecuación de las probabilidades estimadas, cuanto más se acerca el $p - valor$ a cero, peor es la estimación. Sin embargo, si las probabilidades de incumplimiento estimadas son muy pequeñas, la tasa de la convergencia a la distribución χ^2 puede ser muy baja también. Por otra parte, los $p - valores$ proporcionan una posible forma de comparar directamente los pronósticos con diferentes números de categorías de calificación.

La aplicación de esta prueba de Hosmer – Lemeshow se basa en las hipótesis de independencia y de aproximación Normal, por lo que hay que tener en cuenta que posiblemente subestima el verdadero error tipo I. Para un número pequeño de pacientes en cada categoría de calificación, la hipótesis nula es más difícil de rechazar.

7.7.3. Test de Spiegelhalter

Normalmente se calculan individualmente las probabilidades de *cardiohealth default* pronosticadas por el modelo para cada paciente. Puesto que el test Chi – cuadrado de Hosmer – Lemeshow, al igual que el test Binomial, requieren que todos los pacientes asignados a una categoría de calificación tengan la misma probabilidad de *cardiohealth default* es necesario promediar las probabilidades de *cardiohealth default* pronosticadas de los pacientes que han sido clasificados en la misma categoría de calificación, por lo que en los cálculos se pueden introducir algunos sesgos. Este problema puede ser evitado usando el test de Spiegelhalter, que permite las variaciones probabilidades de *default* dentro de la misma categoría de calificación.

Si consideramos N pacientes en el sistema de cuantificación del riesgo y tal que para el paciente $i = 1, \dots, N$, sea c_i la puntuación que el modelo le ha asignado y $\hat{p}_i$ la estimación de la probabilidad de *cardiohealth default* al comienzo de un periodo, observándose al final del periodo el *cardiohealth default* $(y_i = 1)$ o el no *default* $(y_i = 0)$ para cada paciente, el Error Cuadrático Medio *MSE*, viene dado por:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2 \qquad (7.30)$$

El estadístico *MSE* constituye el punto de partida del *Test de Spiegelhalter*.

El estimador $\hat{p}_i$ minimiza el *MSE* esperado si se da un adecuado pronóstico de la probabilidad de cardiohealth *default*. La hipótesis nula para el contraste es que todas las probabilidades de *cardiohealth default* estimadas $\hat{p}_i$, coinciden exactamente con la verdadera aunque desconocida probabilidad de *default* $P(y_i = 1|X = x_i)$ para todo i.

$$H_0: \hat{p}_i = P(y_i = 1|X = x_i) = E[Y|X = x_i], \quad i = 1, \dots N \qquad (7.31)$$

Asumiendo la hipótesis de independencia y usando el Teorema Central del Límite, se demuestra que bajo la hipótesis nula el test estadístico,

$$Z_S = \frac{MSE - E[MSE]}{\sqrt{Var[MSE]}} = \frac{\sum_{i=1}^{N}\{(y_i - \hat{p}_i)^2 - \hat{p}_i(1 - \hat{p}_i)\}}{\sqrt{\sum_{i=1}^{N}\hat{p}_i(1 - \hat{p}_i)(1 - 2\hat{p}_i)^2}}$$

se aproxima a una distribución normal estándar $N(0,1)$, que permite un test de decisión estándar.

CONCLUSIONES

La exhaustiva revisión bibliográfica que se ha realizado sobre los sistemas de predicción del riesgo cardiovascular pone de manifiesto:

(1) Las Tablas de Riesgo Cardiovascular utilizadas en España son:
- *Framingham Regicor*, que es la versión calibrada del estudio norteamericano *Framingham Heart Study.*
- *European Heart Score* en su versión de países de bajo riesgo y versión calibrada.

(2) *Framingham Regicor* calibra el Modelo de Framingham por Categorías de Wilson. Está basado en el Modelo de Riesgos Proporcionales de Cox. Considera el efecto independiente de los factores de riesgo. Supone efectos constantes a lo largo del tiempo. Predice el riesgo coronario a diez años. En la misma línea se plantea el reciente estudio ERICE – Score.

(3) La calibración de *Framingham Regicor* se hace a partir del registro poblacional de infarto de miocardio de Girona REGICOR. Aunque realmente los valores de tasa de incidencia de IAM silente y de angina eran desconocidas, y se asumió que la proporción era similar a la de Framingham. La incidencia de IAM en Girona se encuentra aproximadamente un 15% por debajo del promedio de España.

(4) *European Heart Score* parte de la información de doce estudios de cohorte europeos incluyendo España, pero en realidad vuelven a ser datos de Cataluña. Predice el riesgo cardiovascular total a diez años mediante el Modelo de Riesgos Proporcionales Weibull. Presenta las mismas deficiencias que el modelo anterior. La versión calibrada a partir del estudio MONICA – Catalunya produce riesgos superiores en un 13% al de la función de bajo riesgo.

A partir del escenario actual de estimación del riesgo descrito, se diseña un sistema de cuantificación del riesgo cardiovascular al que denominamos ***Spanish Cardiovascular Risk Scorecard***. El novedoso planteamiento de este sistema de predicción del riesgo resuelve las limitaciones metodológicas esenciales de los sistemas actuales de valoración.

(5) Las predicciones se realizan a partir de Modelos Logísticos Lineales por expansiones lineales Híbridas de funciones de base. La propuesta de funciones

de base para la estimación del riesgo cardiovascular se basa en las funciones constantes a trozos, los pesos de la evidencia y los splines cúbicos.

(6) A partir de este Modelo Logístico Lineal Híbrido es posible calcular la probabilidad de *cardiohealth default*, estimar la función de calificación de pacientes a través del *logit* de la probabilidad de *cardiohealth default a posteriori* y clasificar nuevos pacientes.

(7) Las predicciones se realizan para un horizonte temporal próximo de un año. Se plantea el calibrado del modelo al finalizar el horizonte temporal. Planteamiento absolutamente novedoso en los sistemas de cuantificación del riesgo cardiovascular.

(8) Se diseña el *Algoritmo de Construcción del Spanish Cardiovascular Risk Scorecard*.

Líneas de investigación futuras

(1) La investigación al respecto de las funciones de base óptimas y la elaboración de un detallado *Diccionario de Funciones de Base*, acompañado de algún método para controlar la complejidad del modelo, supone un campo de exploración abierto que conformaría una valiosa herramienta de aplicación real en la estimación del riesgo en las distintas áreas.

(2) La aplicación de la propuesta *Spanish Cardiovascular Risk Scorecard* a una cohorte poblacional española con la colaboración de expertos en el campo de la cardiología.

(3) Una posible mejora podría ser el plantear criterios óptimos de asignación de categorías de calificación. Este es un campo abierto de investigación que deberá ser supervisado tanto por estadísticos como por expertos en riesgo cardiovascular.

Bibliografía

1. Ruiz JS. Epidemiología de la enfermedad cardiovascular: Control global del riesgo cardiometabólico. Diaz de Santos; 2012.

2. Dégano IR, Elosua R, Marrugat J. Epidemiología del síndrome coronario agudo en España: estimación del número de casos y la tendencia de 2005 a 2049. Rev Española Cardiol. 2013;66(6):472–81.

3. Elosua R. Las funciones de riesgo cardiovascular: utilidades y limitaciones. Rev Esp Cardiol. 2014;67(2):77–9.

4. Assmann, G. Paul Cullen HS. Simple Scoring Scheme for Calculating the Risk of Acute Coronary Events Based on the 10-Year Follow-Up of the Prospective Cardiovascular Munster (PROCAM) Study. Circulation [Internet]. 2002 [cited 2014 Jan 8];105:310–5. Available from: http://circ.ahajournals.org/content/105/3/310.short

5. Gabriel R, Brotons C, Tormo MJ, Segura A, Rigo F, Elosua R, et al. The ERICE-score: the New Native Cardiovascular Score for the Low-risk and Aged Mediterranean Population of Spain. Rev Española Cardiol (English Ed [Internet]. 2015 Mar [cited 2015 Mar 11];68(3):205–15. Available from: http://www.sciencedirect.com/science/article/pii/S1885585714002448

6. Dawber TR, Kannel WB, Lye LP. An approach to Longitudinal Studies in a Community: The Framingham Study. Ann N Y Acad Sci. 1963;(107):539–56.

7. Dawber TR, Meadors GF, Moore FE. Epidemiological Disease : The Approaches to Heart Framingham Study. Am J Public Heal Nations Heal. 1951;41(3):279–81.

8. O'Donnell CJ, Elosua R. Cardiovascular Risk Factors. Insights From Framingham Heart Study. Rev Española Cardiol (English Ed [Internet]. Elsevier; 2008 [cited 2015 Apr 22];61(3):299–310. Available from: http://www.revespcardiol.org/en/factores-riesgo-cardiovascular-perspectivas-derivadas/articulo/13117552/

9. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. Circulation [Internet]. 1991 [cited 2014 Oct 28];83:356–62. Available from: http://circ.ahajournals.org/cgi/doi/10.1161/01.CIR.83.1.356

10. J. Truett, J. Cornfield WK. A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham. J chron Dis. 1967;20:511–24.

11. Anderson KM. A Nonproportional Hazards Weibull Accelerated Failure Time Regression Model. Biometrics. 1991;47(1):281–8.

12. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. John Wiley and Sons New York. 1980.

13. Wilson PWF, D'Agostino RB, Levy D, Belanger a. M, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation [Internet]. 1998 [cited 2014 Jul 10];97(18):1837–47. Available from: http://circ.ahajournals.org/cgi/doi/10.1161/01.CIR.97.18.1837

14. Cox DR. Regression Models and Life Tables. J R Stat Soc Ser B. 1972;34(2):187–220.

15. Grundy SM, Pasternak R, Greenland P, Smith S, Fuster V. Assessment of Cardiovascular Risk by Use of Multiple-Risk-Factor Assessment Equations. J Am Coll Cardiol. 1999;(34):1348–59.

16. D'Agostino RB, Russell MW, Huse DM, Ellison RC, Silbershatz H, Wilson PW, et al. Primary and subsequent coronary risk appraisal: New results from the Framingham Study. Am Heart J. 2000 Feb;139(2):272–81.

17. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation [Internet]. 2008 [cited 2014 Mar 12];117(6):743–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18212285

18. Pencina MJ, D'Agostino RB, Larson MG, Massaro JM, Vasan RS. Predicting the 30-Year Risk of Cardiovascular Disease: The Framingham Heart Study. Circulation [Internet]. 2009 [cited 2014 Aug 25];119:3078–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2748236&tool=pmcentrez&rendertype=abstract

19. Menotti, A; Lanti, M; Puddu, P E; Kromhout D. Coronary heart disease incidence in northern and southern European populations: a reanalysis of the seven countries study for a European coronary risk chart. Heart [Internet]. 2000 [cited 2014 Oct 28];84(3):238–44. Available from: http://heart.bmj.com/content/84/3/238.short

20. De Backer G, Ambrosionie E, Borch-Johnsen K, Brotons C, Cifkova R, Dallongeville J, et al. European guidelines on cardiovascular disease prevention in clinical practice: Third Joint Task Force of European and other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by

representatives of eight societies and by invit. Eur J Cardiovasc Prev Rehabil [Internet]. 2003 [cited 2013 Nov 17];10(1 Suppl):S1–78. Available from: http://cpr.sagepub.com/content/10/1_suppl/S1.short

21. Marrugat J, Solanas P, D'Agostino R, Sullivan L, Ordovas J, Cordón F, et al. Estimación del riesgo coronario en España mediante la ecuación de Framingham calibrada. Rev Española Cardiol. 2003;56(3):253–61.

22. Pérez G, Pena A, Sala J, Roset P, Masiá R, Marrugat J and the RI. Acute myocardial infarction case fatality, incidence and mortality rates in a population registry in Gerona, Spain, 1990-1992. Int J Epidemiol. 1998;27:599–604.

23. Tablas de evaluación del riesgo coronario adaptadas a la población española. Estudio DORICA. Med Clin (Barc). 2004;123(18):686–91.

24. Marrugat J, D'Agostino R, Sullivan L, Elosua R, Wilson P, Ordovas J, et al. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. J Epidemiol Community Health. 2003 Aug;57:634–8.

25. Marrugat J, Fiol M, Sala J, Tormo M. Variabilidad geográfica en España en las tasas de incidencia y mortalidad poblacionales por infarto agudo de miocardio en el estudio IBERICA. Rev Esp Cardiol. 2000;14(2):81.

26. Cristóbal J, Lago F, de la Fuente J, González-Juanatey JR, Vázquez-Bellés P, Vila M. Ecuación de Framingham de Wilson y ecuación de REGICOR. Estudio comparativo. Rev Española Cardiol. 2005;58(8):910–5.

27. Baena-Díez JM, Ramos R, Marrugat J. Capacidad predictiva de las funciones de riesgo cardiovascular: limitaciones y oportunidades. Rev Española Cardiol Supl. 2009;9:4–13.

28. Assmann G, Schulte H, Cullen P, Seedorf U. Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Münster (PROCAM) study. Eur J Clin Invest [Internet]. 2007 [cited 2014 Apr 29];37(12):925–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18036028

29. D'Agostino, Sr RB, Grundy S, Sullivan LM, Wilson P. Validation of the Framingham Coronary Heart Disease Prediction Scores. JAMA [Internet]. American Medical Association; 2001 [cited 2014 Sep 5];286(2):180–7. Available from: http://jama.jamanetwork.com/article.aspx?articleid=193997

30. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, Backer G. DP.

Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24:987–1003.

31. Mancia G, Fagard R, Narkiewicz K, Redón J, Zanchetti A, Böhm M, et al. 2013 European Society of Hipertension-European Society of Cardiology guidelines for the management of arterial hypertension. J Hypertens [Internet]. 2013 [cited 2014 Mar 20];31(7):1281–357. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23817082

32. Kesteloot H. Dynamics of cardiovascular and all-cause mortality in Western and Eastern Europe between 1970 and 2000. Eur Heart J [Internet]. 2006;27:107–13. Available from: http://eurheartj.oxfordjournals.org/cgi/doi/10.1093/eurheartj/ehi511

33. Sans S, Fitzgerald AP, Royo D, Conroy R, Graham I. Calibración de la tabla SCORE de riesgo cardiovascular para España. Rev Española Cardiol. 2007;60(5):476–85.

34. Cooney MT, Dudina A, D'Agostino R, Graham IM. Cardiovascular risk-estimation systems in primary prevention: do they differ? Do they make a difference? Can we see the future? Circulation [Internet]. 2010 [cited 2013 Nov 27];122(3):300–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20644026

35. Bordonada R, Angel M, Sala Á, De F, De A, Ma J, et al. Adaptación española de la guía europea de prevención cardiovascular. Rev Esp Salud Pública. 2004;78(4):435–8.

36. García Mora R, Félix Redondo FJ. Concordancia de dos métodos para el cálculo del riesgo cardiovascular: Framingham calibrado por REGICOR y SCORE. Hipertensión. 2005;22(8):306–10.

37. Buitrago Ramírez F, Cañón Barroso L, Díaz Herrera N, Cruces Muro E, Bravo Simón B, Pérez Sánchez I. Comparación entre la tabla del SCORE y la función Framingham-REGICOR en la estimación del riesgo cardiovascular en una población urbana seguida durante 10 años. Med Clin (Barc). 2006;127(10):368–73.

38. Brotons C, Cascant P, Ribera A, Moral I, Permanyer G. Utilidad de la medición del riesgo coronario a partir de la ecuación del estudio de Framingham: estudio de casos y controles. Med Clin (Barc). 2003;121(9):327–30.

39. May S, Hosmer DW. Advances in Survival Analysis. Handbook of Statistics.

Elsevier; 2004. 383-394 p.

40. DiRienzo AG, Lagakos SW. Advances in Survival Analysis. Handbook of Statistics. Elsevier; 2004. 395-409 p.

41. Bagdonavičius V, Nikulin M. Advances in Survival Analysis. Handbook of Statistics. Elsevier; 2004. 411-429 p.

42. Marrugat J, Subirana I, Comín E, Cabezas C, Vila J, Elosua R, et al. Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA study. J Epidemiol Community Health [Internet]. 2007 Jan [cited 2014 Apr 10];61:40–7. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2465597&tool=pmcentrez&rendertype=abstract

43. Galve E, Castro A, Cordero A, Dalmau R, Fácila L, García-Romero A, et al. Update in cardiology: vascular risk and cardiac rehabilitation. Rev española Cardiol [Internet]. 2013 [cited 2014 Aug 25];66(2):124–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23266065

44. Rodríguez-Artalejo F, Banegas JB. De la ecuación de Framingham a la prevención cardiovascular. Med Clin (Barc). 2003;121(9):334–6.

45. Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol [Internet]. Elsevier Inc.; 2009 [cited 2014 Jul 9];54(14):1209–27. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19778661

46. Brotons C, Moral I, Soriano N, Cuixart L, Osorio D, Bottaro D, et al. Impacto de la utilización de las diferentes tablas SCORE en el cálculo del riesgo cardiovascular. Rev Esp Cardiol [Internet]. 2014 [cited 2014 Mar 20];67(2):94–100. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24268126

47. Maiques A, Antón F, Franch M, Albert X, Aleixandre E, Gil C, et al. Riesgo cardiovascular del SCORE comparado con el de Framingham . Consecuencias del cambio propuesto por las Sociedades Europeas. Med Clin (Barc). 2004;18(123):681–5.

48. Marrugat J, Vila J, Baena-Díez JM, Grau M, Sala J, Ramos R, et al. Relative Validity of the 10-Year Cardiovascular Risk Estimate in a Population Cohort of the REGICOR Study. Rev española Cardiol [Internet]. 2011 [cited 2014 Aug 25];64(5):385–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21482004

49. Perk J, De Backer G, Gohlke H, Graham I, Reiner Z, Verschuren M, et al. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). Eur Heart J [Internet]. 2012 [cited 2014 Jul 10];33(13):1635–701. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22555213

50. Cooney MT, Vartiainen E, Laatikainen T, De Bacquer D, McGorrian C, Dudina A, et al. Cardiovascular risk age: concepts and practicalities. Heart [Internet]. 2012 [cited 2014 Oct 1];98:941–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22626902

51. Wilkins JT, Ning H, Berry J, Zhao L, Dyer AR, Lloyd-Jones DM. Lifetime Risk and Years Lived Free of Total Cardiovascular Disease. JAMA [Internet]. 2012;308(17):1795–801. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3748966&tool=pmcentrez&rendertype=abstract

52. Brotons C, Lobos JM, Royo-Bordonada MÁ, Maiques A, de Santiago A, Castellanos Á, et al. Implementation of Spanish adaptation of the European guidelines on cardiovascular disease prevention in primary care. BMC Fam Pract [Internet]. 2013;(14):36. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3608081&tool=pmcentrez&rendertype=abstract

53. Mallo Fernández F. Modelos Multivariantes Internos de Medición de Riesgos de Crédito, Acordes con Basilea II. Tesis doctoral Universidad de Salamanca. 2011.

54. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen. 1936;7:179–88.

55. Ladd GW. Linear probability functions and discriminant function. Econometrica. 1966;34:873–855.

56. Lachenbruch PA. Discriminant Analysis. Press H, editor. New York; 1975.

57. Hastie, T. and Tibshirani R. Nonparametric regression and classification. From Stat to Nerual Networks Theory Pattern Recognit Appl Comput Sci. 1996;136:78–82.

58. Belsey, D. A., Kuh, E., Welsch RE. Regression Diagnostics: Identifying influential data and sources of collinearity. John Wiley, New York; 1980.

59. Kleinbaum, D. G., Kupper, L.L., Muller KE. Applied regression analysis and other multivariate methods. PWS-Kent, Boston; 1988.

60. Kullback S. Information Theory and statistics. Wiley, New York; 1959.

61. Hastie, T., Tibshirani, R., Friedmann J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2ª Edition. 2009.

62. Siddiqi N. Credit Risk Scorecards. Developing and implementing Intelligent Credit Scoring. Hoboken, New Yersey: John Wiley & Sons, Inc.; 2006.

63. Breiman, L., Friedman, J.H., Olshen, R.A., Stone CJ. Classification and Regression Trees. Belmont, Wadsworth: Chapman & Hall; 1984.

64. Stone, C.J., Koo CY. Additive splines in statistics. Proc Stat Comput Sect ASA, Am Stat Assoc. 1985;45–8.

65. McFadden D. Conditional logit analysis of qualitative choice behaviour. Zaremba, editor. Frontiers in Econometrics. New York: Academic Press; 1974. 105-142 p.

66. Cox, D.R., Snell EJ. Analysis of Binary data. London: Chapman & Hall; 1989.

67. Nagelkerke NJD. A note on a General Definition of the Coeficient of Determination. Biometrika. 1991;(78):691–2.