71/2003-04

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
UNIDAD DE TERCER CICLO Y POSTGRADO

Reunido el día de la fecha, el Tribunal nombrado por el Excmo. Sr. Rector Magfco. de esta Universidad, el/a aspirante expuso esta TESIS DOCTORAL.
Terminada la lectura y contestadas por el/a Doctorando/a las objeciones formuladas por los señores miembros del Tribunal, éste calificó dicho trabajo con la nota de SOBRESALIENTE CUM LAUDE POR UNANIMIDAD.

Las Palmas de Gran Canaria, a 17 de septiembre de 2004.

El/la Presidente/a: Dr. D. Miguel Sánchez García

El/la Secretario/a: Dr. D. José María Limiñana Cañal

El/la Vocal: Dr. D. María Inés Sobrón Fernández

El/la Vocal: Dr. D. Angelo Santana del Pino

El/la Vocal: Dr. D. Pedro Moreno Ramis

El/la Doctorando/a: D. Eduardo Núñez González

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

DEPARTAMENTO DE MATEMÁTICAS.

ANÁLISIS DE FACTORES DE RIESGO EN HIPERTENSIÓN ARTERIAL. UNA VISIÓN MATEMÁTICA.

EDUARDO NÚÑEZ GONZÁLEZ.

2004

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

DIRECTOR: DR. PEDRO SAAVEDRA SANTANA

CODIRECTOR: DR. DOMINGO GUZMÁN PÉREZ HERNÁNDEZ

ANÁLISIS DE FACTORES DE RIESGO EN HIPERTENSIÓN ARTERIAL. UNA VISIÓN MATEMÁTICA.

EDUARDO NÚÑEZ GONZÁLEZ.

PEDRO SAAVEDRA SANTANA, Catedrático de Universidad del Área de Conocimiento de Estadística e Investigación Operativa del Departamento de Matemáticas de la Universidad de Las Palmas de Gran Canaria.

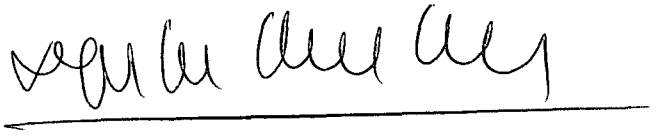
CERTIFICA: Que la presente memoria titulada **Análisis de factores de riesgo en hipertensión arterial. Una visión matemática,** ha sido realizada bajo mi dirección por el Licenciado en Ciencias Matemáticas D. Eduardo Núñez González, y constituye su Tesis para optar al grado de Doctor en Matemáticas.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos a los que de lugar, firmo la presente en Las Palmas de Gran Canaria, a treinta de junio de dos mil cuatro.

DOMINGO GUZMAN PÉREZ HERNÁNDEZ, médico especialista en Geriatría, profesor asociado del Departamento de Enfermería de la Universidad de Las Palmas de Gran Canaria y Director del Hospital Insular de Lanzarote.

CERTIFICA: Que la presente memoria titulada **Análisis de factores de riesgo en hipertensión arterial. Una visión matemática,** ha sido realizada bajo mi dirección por el Licenciado en Ciencias Matemáticas D. Eduardo Núñez González, y constituye su Tesis para optar al grado de Doctor en Matemáticas.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos a los que de lugar, firmo la presente en Las Palmas de Gran Canaria, a treinta de junio de dos mil cuatro.



A MI FAMILIA,

POR EL TIEMPO INEVITABLEMENTE ROBADO.

Mi agradecimiento al profesor Dr. D. Pedro Saavedra Santana, director de esta memoria, por todo.

Mi reconocimiento al Dr. D. Domingo Guzmán Pérez Hernández, codirector de esta memoria, por su ayuda.

Capítulo II: Métodos Estadísticos.

Capítulo III: Estudio de hipertensión arterial en Lanzarote.

Capítulo IV: Discusión y Conclusiones.

Prólogo.

Existe una teoría que considera que la Matemática se puede contemplar como conjunto de ciencias, incluso alguna de ellas natural. Sin embargo, uno de los hallazgos culturales decisivos del siglo XIX fue el descubrimiento que la Matemática no es una ciencia natural, sino una creación intelectual del hombre. Aún aceptando esta segunda acepción, la creación suele nacer de la necesidad de responder a las demandas de la ciencia y de la técnica, consideradas estas en sentido amplio.

Una concreción, en resultados tangibles a estas demandas científicas, son las aplicaciones al mundo de Ciencias de la Salud o Ciencias de la Vida. La Estadística, como parte de la Matemática, y su contextualización en las Ciencias de la Vida, la Bioestadística, constituyen un claro ejemplo de cómo la Matemática Aplicada incide en un campo de estudio e investigación revelándose como un instrumento imprescindible, que nos ayuda a descubrir causas, detectar y analizar características comunes, valorar factores; en definitiva, aproximarnos a lo verdadero.

Basta hojear publicaciones como *Lancet o British Medical Journal*, para darnos cuenta de que la mayoría de estudios contienen análisis o valoraciones estadísticas. No parece fuera de lugar pensar que parte del gran desarrollo habido en la Matemática se debe a la demanda procedente de otras ciencias.

Por otro lado, se han creado y popularizado vehículos de aplicación muy importantes. El ordenador como plataforma y los paquetes estadísticos como aplicación han puesto esta herramienta al alcance de todos. No obstante, conviene recordar que la informática solamente desarrolla métodos y obtiene resultados, no los planifica ni los interpreta; cuestiones que constituyen una fuente de error importante. La computación también apoya la investigación de un modo fundamental: uno de los métodos desarrollados en esta memoria, *bootstrap*, no se puede llevar a cabo sin su ayuda. Pero los mejores frutos de este maridaje, Estadística, Estocástica y Computación aún están por llegar; los sistemas expertos para ayuda al diagnóstico médico, la imagen computerizada y la computación neuronal, son campos en los que apenas hemos conseguido rascar la superficie.

La presente memoria tiene un marcado carácter metodológico, se desarrollan métodos ya conocidos tanto clásicos como punteros, aclarando algunos de sus extremos, esclareciendo en que circunstancias pueden aplicarse y, en definitiva, puliendo y, en cierto modo, validando las diversas metodologías empleadas. Se trata de un estudio epidemiológico sobre hipertensión arterial hecho en Lanzarote. En concreto se realiza un análisis sobre los factores de riesgo asociados a esta enfermedad. El diseño del estudio es de caso-control. Apoyándonos en unos datos, trata de determinar factores de riesgo para esta enfermedad; también se estudian los efectos de estas variables, en concreto de la hipertensión, en distintos tipos de enfermedades.

El trabajo esta organizado en cuatro capítulos. En el primero se plantean los objetivos, se explica la etiología de la hipertensión arterial, los factores que influyen ella y sus consecuencias. Se revisa brevemente el conocimiento existente sobre esta enfermedad.

El segundo capítulo contiene la exposición de los métodos estadísticos. Se desarrollan y comentan test de hipótesis paramétricos como el test de la *t de Student*, conviviendo con otras metodologías de libre distribución, como test de las *permutaciones bootstrap* o el de *Wilcoxon.* Plantea y desarrolla la familia de transformaciones de *Box y Cox,* para conseguir normalidad conjunta para casos y controles. Se trata el *confunding* y se apuntan posibles soluciones. El estudio de posibles marcadores para el diagnóstico se desarrolla utilizando curvas *ROC* y *análisis discriminante*. Los modelos de *regresión logística* ocupan lugar importante en esta memoria, pues en primer lugar se desarrollan y, posteriormente, se aportan soluciones al problema de la inflación del vector de coeficientes cuando existe escasez de datos. El *bootstrap* es utilizado profusamente en este trabajo.

En el tercer capítulo, se exponen los resultados de un estudio de hipertensión arterial en Lanzarote. Se aplican todas las metodologías expuestas en el capítulo anterior, tratando de determinar factores asociados a hipertensión arterial, que supongan riesgo para los individuos expuestos. Contiene un interesante estudio de medias. Se intenta ir más allá de la propia enfermedad, y se estudian posibles repercusiones sobre otros órganos, como corazón o cerebro.

En el cuarto y último capítulo se discuten los resultados del estudio, en la doble vertiente médica y estadística, planteadas en esta memoria como partes de un todo. Se obtienen conclusiones. De alguna manera, también se valoran las propias metodologías empleadas.

Auque ya hemos mencionada la importancia que las aplicaciones de la computación tienen para Ciencias de la Vida; conviene resaltar que en este trabajo se han utilizado, además de paquetes estándar como *SPSS* o *R*, programación a medida. Así, se han desarrolla aplicaciones en *Pascal* o *C++*, para implementar métodos no contenidos en los paquetes. En el CD-Rom que acompaña a esta memoria, se encuentran tanto estos programas, cuyos listados figuran en el segundo apéndice de este trabajo, como las bases de datos utilizadas.

Resulta difícil concluir esta introducción sin hablar específicamente del análisis de datos. Son cada vez más las peticiones en este sentido, que llegan de mundos tan dispares como economía, industria, biología o medicina. Dar respuesta no siempre es sencillo, se requiere un conocimiento del problema planteado, un lenguaje común con el demandante y fuertes conocimientos de, entre otras materias, metodología de la investigación, álgebra, análisis, estadística y computación. Dentro del arte que conlleva la Matemática, en ocasiones se tiende a sublimarse; a tener un pensamiento totalmente abstracto, huyendo de lo concreto. Este hecho necesario e incluso imprescindible, no debería llevar al olvido de la Matemática Aplicada, pues son las demandas procedentes de otras ciencias y la propia realidad social, quienes, en muchas ocasiones, actúan como motor de desarrollo, colaborando a añadir pisos a este gran edificio que es la Matemática.

Capítulo I: Hipertensión Arterial.

1.1. Objetivos.

En la presente memoria se desarrolla una metodología estadística moderna orientada consiste en un estudio sobre hipertensión arterial, enfocado desde el doble punto de vista clínico y matemático. Se desarrollan diversas metodologías estadísticas de vanguardia, ejemplificadas en unos datos concretos, que son puntuaciones de algunas de las variables descritas al principio del tercer capítulo. Se traza como objetivos:

- Buscar factores asociados a la enfermedad (de riesgo), que sean causa de la misma; entendiendo como factor de riesgo aquella característica biológica o social que hipotéticamente podría dar lugar a hipertensión arterial.

- Analizar protocolos diagnósticos, en relación tanto con su bondad como con su eficacia desde distintos puntos de vista.

- Desarrollar, pulir y concretar métodos estadísticos de vanguardia que si, bien son conocidos, necesitan de una mejor plástica o profundización, concretándolos en algunos aspectos, relacionándolos entre sí, o desarrollando herramientas que posibiliten su uso y mejor comprensión.

Se presta una especial atención al tratamiento previo de datos y a detectar posibles factores de confusión.

1.2. Hipertensión arterial. Estado del Arte.

1.2.1. Introducción.

La hipertensión arterial es uno de los factores de riesgo más directamente relacionados con la mortalidad cardiovascular. Se trata, por tanto, de uno de los problemas de salud pública más importantes de los países desarrollados. La hipertensión arterial no tratada predispone al desarrollo de cardiopatía isquémica, insuficiencia cardiaca, ictus, insuficiencia renal y enfermedad vascular periférica. Aunque se han hecho grandes avances en el conocimiento de la fisiopatología, en el tratamiento de la hipertensión y en la prevención de las complicaciones, la etiología de la enfermedad sigue en gran parte sin conocerse. Los progresos realizados en las últimas décadas sobre la reducción de la mortalidad cardiovascular están directamente relacionados con los programas de control de la hipertensión arterial y de otros factores de riesgo cardiovascular. Pero el desconocimiento que aún se tiene de la etiología de la enfermedad dificulta un tratamiento más específico e individualizado, lo que facilita el incumplimiento terapéutico y, por tanto, limita los esfuerzos destinados a conseguir una mayor reducción de la mortalidad por enfermedades cardiovasculares.

1.2.2. Definición.

Se entiende por hipertensión arterial la elevación persistente de la presión arterial sistólica o diastólica. La comparación de distintos estudios epidemiológicos muestra que las presiones sanguíneas siguen, en cualquier población, una distribución aproximadamente normal. Por tanto los límites a partir de los cuales se considera a las cifras de tensión arterial como

hipertensión, se definen de forma arbitraria y representan el límite superior de una variable continua. Aunque el consenso sobre las cifras a partir de las cuales se puede hablar de hipertensión ha ido variando a lo largo del tiempo, hoy se acepta que este concepto puede definirse como *una presión arterial sistólica mayor o igual a 140 mm Hg. o una presión diastólica mayor o igual a 90*, con independencia de la edad (Criterio de la OMS 1993).

1.2.3. Prevalencia.

La prevalencia de la hipertensión arterial varía en función de la población estudiada y de los criterios empleados para definirla. De ahí que sea difícil comparar los distintos estudios epidemiológicos, debido a la variación en las distintas cifras utilizadas para definir la hipertensión arterial y, sobre todo, a que las características de las poblaciones estudiadas no siempre son comparables. La estimación de la prevalencia de hipertensión en una población suburbana de raza blanca como la utilizada en el estudio Framingham, dio como resultado que una quinta parte de la población tenía presiones arteriales superiores a 160/95 y casi la mitad tenía presiones superiores a 140/90.

Entre las personas que viven en los países industrializados, las presiones arteriales sistólicas y diastólicas suelen aumentar hasta, aproximadamente, los sesenta años. La prevalencia en las mujeres se relaciona con la edad y aumenta a partir de los cincuenta años, probablemente ligado a los cambios hormonales de la menopausia. Sin embargo, en los países subdesarrollados ni la presión sistólica ni la diastólica aumentan con la edad, y la hipertensión es casi inexistente,

debido posiblemente a la baja ingesta de sodio y al mayor nivel de actividad física.

Grandes estudios prospectivos, bien diseñados, desde la década de los setenta, han demostrado la fuerte asociación entre la hipertensión arterial y las enfermedades cardiovasculares, en particular la cardiopatía isquémica. La prevalencia de la hipertensión entre la población adulta en España se sitúa en el 34.2% para la población comprendida entre los 35 y los 64 años. En cuanto a los datos disponibles en la población de las Islas Canarias, Dorta y colaboradores encuentran una prevalencia de la hipertensión arterial en la población de ambos sexos en la isla de Tenerife del 19% para la comprendida entre los 17 y los 74 años).
Rodríguez Pérez, en un estudio realizado en Lanzarote entre la población de ambos sexos, en edades situadas entre los 30 y los 64 años, encuentra una prevalencia de hipertensión arterial del 26%.

1.2.4. Etiología.

La hipertensión arterial *esencial*, primaria o idiopática, sería aquella que no tiene una causa definible. La variedad de mecanismos que regulan la presión arterial -adrenérgicos, periféricos o centrales, renales, hormonales y vasculares- y sus complejas relaciones, probablemente dificulten identificar el origen de la hipertensión arterial en estos pacientes. Se han descrito diferentes alteraciones en los pacientes con hipertensión arterial esencial. No se sabe si estas alteraciones son expresiones variables de un mismo proceso, aunque lo más probable es que se trate del reflejo de distintas entidades independientes. Por tanto la hipertensión esencial probablemente se deba a diversas causas. El 90% de

los adultos que padecen hipertensión, estarían incluidos dentro de este grupo.

La hipertensión *secundaria* es menos frecuente que la esencial y puede deberse a diversos trastornos que también pueden agravar una hipertensión esencial previa. Los más frecuentes, entre estos trastornos son las enfermedades renales, incluyendo la estenosis arterial renal por arteriosclerosis, las neoplasias renales, la nefropatía parenquimatosa por glomerulonefritis crónica, pielonefritis crónica, poliquistosis renal, nefropatía de las enfermedades del colágeno y uropatía obstructiva.

Los trastornos endocrinos pueden causar hipertensión y son a veces responsables de la aparición de hipertensión de comienzo reciente en los ancianos. En este grupo podríamos encuadrar a la hipertensión secundaria, enfermedades tiroideas, síndrome de Cushing, aldosteronismo primario, estados hipercalcémicos y liberación de hormonas o sustancias que se comportan como tales a partir de tumores malignos.

Muchos fármacos pueden provocar hipertensión y agravar o complicar una hipertensión previa. Entre los primeros, entre otros, estarían: la ciclosporina, los antidepresivos tricíclicos, los inhibidores de la monoamino oxidasa, la fenilpropanolamina y otros vasoconstrictores presentes en preparados farmacéuticos de venta libre; los corticoides y los estrógenos. Entre los fármacos que pueden agravar la hipertensión habría que destacar a los antiinflamatorios no esteroides. La hipertensión arterial también puede deberse al consumo de cocaína, regaliz o exceso de alcohol, así como a coartación aórtica.

La enfermedad arteriosclerótica puede dar lugar a hipertensión *sistólica aislada*, particularmente frecuente en los ancianos, al reducir la distensibilidad de las grandes arterias. Como el ventrículo debe impulsar su volumen sistólico hacia una aorta menos distensible, hace que aumente la poscarga de presión sistólica. Otros procesos que pueden contribuir a la presión sistólica aislada en los ancianos serían: el hipertiroidismo, la insuficiencia aórtica, las desnutrición, las fístulas arteriovenosas y la fiebre.

1.2.5. Factores etiológicos.

1.2.5.1. Hipertensión y edad.

Las presiones arteriales sistólica y diastólica suelen aumentar hasta alrededor de los sesenta años. La presión arterial sistólica puede seguir aumentando a partir de entonces, pero la diastólica tiende a estabilizarse o disminuir. Esto en los que se refiere a los países industrializados, donde hasta el 50% de los mayores de 65 años pueden cumplir con las características descritas, porque los habitantes de algunos países subdesarrollados, ni la presión sistólica ni la diastólica aumentan con la edad. Por otro lado, la alta prevalencia de la hipertensión, podría sugerir que el aumento de la presión sistólica relacionada con la edad es normal e incluso inocuo, si bien existen estudios que desmienten esto último. Sin embargo, cuanto más elevadas sean la presión arterial sistólica o la diastólica, más elevada será la morbi-mortalidad total de origen cardiovascular.

Las características hemodinámicas de los hipertensos ancianos, son similares a las de los jóvenes. Sin embargo, en los ancianos, la resistencia

periférica total calculada puede ser más elevada y menos la compliancia de las grandes arterias. En la mayoría de los ancianos con hipertensión arterial esencial, el volumen intravascular se contrae en la medida que aumenta la presión arterial y la resistencia periférica total. El estudio de Framingham demuestra que el impacto de la presión sistólica es similar al de la diastólica. Incluso con hipertensión arterial sistólica asilada se produce un aumento significativo del riesgo cardiovascular, y este no declina en edades avanzadas.

1.2.5.2. Hipertensión y herencia.

Datos de numerosos estudios han apoyado la importancia de los factores genéticos en el origen de la hipertensión arterial. Uno de los puntos de vista ha sido la agregación familiar, o correlación de hipertensos dentro las mismas familias. Sin embargo, la mayor parte de los estudios apoya el concepto de que la herencia es probablemente multifactorial, o que diversos defectos genéticos diferentes tienen como expresión fenotípica la elevación de la presión arterial. Se ha encontrado una recurrencia familiar de la hipertensión entre parientes de primer grado. Asimismo, los niveles de presión arterial en niños naturales se correlacionan más directamente con los de sus progenitores que los niveles de presión de los niños adoptados con los de sus padres adoptivos.

1.2.5.3. Hipertensión y ambiente.

Numerosos factores ambientales han sido relacionados con la hipertensión arterial. Entre los más conocidos se encuentran el consumo de sal, la obesidad, la profesión, el consumo de alcohol, el tamaño de la familia y el hacinamiento. La presencia de alguno o varios de estos factores explicarían la alta prevalencia de la hipertensión arterial en las

sociedades prósperas, y su elevación con la edad, al contrario de lo que ocurre en los países menos desarrollados. Está fuera de duda la asociación entre hipertensión y exceso de peso. Los aumentos de peso conducen a incrementos de la presión arterial, tanto en hombres como en mujeres, y a cualquier edad. Según el estudio de Framingham, cada unidad de incremento en el índice de masa corporal, se asocia a un aumento mínimo de 1 mm Hg en la presión sistólica.

1.2.5.4. Hipertensión y consumo de sal.

Es uno de los factores ambientales más estudiados. Estudios transversales han encontrado una baja prevalencia de hipertensión en aquellas poblaciones con una ingesta de sodio baja. La presión arterial sólo es sensible al consumo de sal en aproximadamente el 60% de los hipertensos, lo que pone de relieve la naturaleza heterogénea de la población que padece hipertensión arterial. La causa de la sensibilidad especial a la sal es variable. En la mitad aproximadamente de los pacientes, la causa habría que buscarla en el aldosteronismo primario, la estenosis bilateral de la arteria renal, las enfermedades parenquimatosas renales, o la hipertensión arterial con renina baja. En el resto la fisiopatología es incierta, pero se han propuesto como factores coadyuvantes el consumo de cloruro y de calcio, un defecto generalizado de las membranas celulares y la resistencia a la insulina.

1.2.5.5. El papel de la renina.

La función endocrina del riñón, mediada por la enzima renina, es un componente importante de la regulación de la homeostasis de la tensión arterial. La primera prueba de la influencia de la secreción anómala de

renina en la enfermedad hipertensiva humana se obtuvo en 1960 de estudios de pacientes con hipertensión maligna. En esta forma grave de hipertensión se observaron excesos sorprendentes de la secreción de aldosterona que en la exploración quirúrgica resultaron deberse a hiperplasia suprarrenal bilateral, no mejorando la hipertensión maligna con la suprarrenalectomía bilateral.

La renina se segrega en las células yuxtaglomerulares del riñón, y está relacionada con la aldosterona a través de un circuito de retroalimentación negativa. Aunque la secreción de renina puede modificarse por diversos factores, el principal es la situación de volumen en el individuo, en especial en lo que se refiere a las variaciones en la ingesta dietética de sodio. El resultado de la acción de la renina sobre sus sustrato es la producción del péptido angiotensina II. La respuesta de los tejidos diana a este péptido, esta determinada por la ingestión previa de electrolitos en la dieta. La ingesta de sodio, en condiciones normales, modula las respuestas vasculares suprarrenales y renales a la angiotensina II. Con la restricción de sodio, las respuestas suprarrenales se facilitan y las vasculares se inhiben y con la sobrecarga de sodio, el efecto es el opuesto. El intervalo de actividad de la renina plasmática que se observa en hipertensos es más amplio que el que se observa en los normotensos. Este hecho ha llevado a clasificar a los pacientes que padecen hipertensión esencial en dos tipos: con renina alta y con renina baja.

1.2.5.5.1 Hipertensión esencial con renina baja.

Un subgrupo de los pacientes que padecen hipertensión arterial esencial, que podría cifrarse en torno al 20%, y que aparece con más frecuencia en la comunidad afroamericana de los EEEUU, presentan una supresión de la actividad de la renina plasmática. Estos pacientes muestran

retención y sodio y una expansión de los volúmenes extracelulares. Se cree, aunque no se ha demostrado aún, que este efecto es debido a la producción excesiva de algún mineralocorticoide no identificado.

1.2.5.5.2. Hipertensión esencial no modulada.

Entre el 25 y 35% de los pacientes con hipertensión arterial esencial tienen niveles de actividad de renina plasmática normal o alta con una dieta pobre en sal, y padecen una forma de hipertensión sensible a la sal debido a un defecto de la capacidad del riñón de excretar adecuadamente el sodio. En estos pacientes, la ingesta de sodio no modula la respuesta suprarrenal, ni la respuesta vascular renal a la angiotensina II. Este subgrupo de hipertensos ha sido denominado no moduladores, debido a la ausencia de modulación de la respuesta de los tejidos diana a la angiotensina II, mediada por el sodio. Al parecer esta anomalía está determinada genéticamente y puede corregirse mediante la administración de inhibidores de la enzima convertidora de la angiotensina.

1.2.5.5.3 Hipertensión esencial con renina alta.

Entorno al 15% de los hipertensos esenciales presentan niveles de actividad de renina plasmática superiores a los valores normales. Se ha sugerido que la actividad de la renina plasmática desempeñaría un papel importante en la patogenia de este tipo de hipertensión. Sin embargo se ha visto que menos de la mitad de estos pacientes responden a tratamientos con antagonistas competitivos de la angiotensina II, lo que ha llevado a algunos autores a proponer que tanto la renina elevada como la presión arterial elevada se deben a una elevada actividad adrenérgica.

1.2.5.5.4. El papel del ion cloruro y del calcio.

La mayor parte de los estudios que han valorado la importancia de la sal en la génesis de la hipertensión arterial han supuesto que el ion sodio es lo esencial. Sin embargo, estudios observacionales en animales hipertensos sensibles a la sal, alimentados con sales sódicas sin contenido de cloruro, no aumentaban su opresión arterial. Por otro lado, en estudios epidemiológicos se ha asociado la ingesta baja de calcio con un aumento de la presión arterial. También es conocido el papel eficaz de los antagonistas del calcio, como antihipertensivos.

1.2.5.5.5. Hipertensión y defecto de la membrana celular.

Estudios que describen alteraciones en el transporte del sodio a través de la membrana celular de los hematíes, entre otros, han dado lugar a la hipótesis del papel de un defecto generalizado de la membrana celular en la génesis de la hipertensión arterial sensible a la sal. Se ha supuesto que esta alteración afectaría a todas las células del organismo, especialmente a las musculares lisas vasculares. Basándose en los estudios sobre hematíes, se ha propuesto que este defecto puede alcanzar desde un 35 a un 50% de los hipertensos esenciales.

1.2.5.5.6. Resistencia a la insulina.

Existe una correlación positiva entre la diabetes y la hipertensión, siendo independiente del índice de masa corporal y de la frecuencia cardiaca basal. Aunque una parte de la población hipertensa presenta resistencia a la insulina e hiperinsulinemia, no se sabe a ciencia cierta si estos hallazgos son los responsables de la elevación de la presión arterial o si se trata de una asociación casual. La resistencia a la insulina es frecuente en los pacientes no insulinodependientes u obesos. Tanto la obesidad como la diabetes mellitus no insulinodependiente son más

frecuentes en los hipertensos que en los normotensos. Por otro lado, se sabe que la hiperinsulinemia es uno de los mecanismos aterogénicos más conocidos en los diabéticos. La hiperinsulinemia puede aumentar la presión arterial por alguno o varios de los siguientes mecanismos. En primer lugar, la hiperinsulinemia produce retención de sodio a nivel renal y aumenta la actividad simpática. Otro mecanismo podría estar relacionado con la hipertrofia del músculo liso vascular secundaria a la acción mitogénica de la insulina. También la insulina modifica el transporte de iones a través de la membrana celular, incrementando así los niveles de calcio en los tejidos vasculares o renales sensibles a la insulina. Sin embargo, el papel de la insulina en el control de la presión arterial solo se conoce de forma vaga y por tanto sigue sin estar claro su papel en la patogenia de la hipertensión arterial.

1.2.6. Consecuencias de la hipertensión arterial.

Los pacientes con hipertensión arterial no complicada, es decir, sin lesiones sobre los órganos diana, suelen permanecer asintomáticos. Aunque la cefalea, la epixtasis y los acúfenos, suelen atribuirse a hipertensión, la frecuencia de estos síntomas no es diferente en los hipertensos que en los normotensos.

En pacientes con hipertensión esencial, la presencia de síntomas o signos sugiere lesión de los órganos diana. Las manifestaciones precoces de afectación cardiaca comprenden cansancio fácil, palpitaciones y ectopia auricular o ventricular.

Los individuos que padecen hipertensión arterial mueren prematuramente. La causa más frecuente es la afectación cardiaca. Según

Brunner (1972), el riesgo de muerte súbita es diez veces superior, cuando además se presentan alteraciones la repolarización. También son causa frecuente el ictus y la insuficiencia renal.

1.2.6.1. Efectos sobre el corazón.

El incremento de la presión arterial sistémica supone una sobrecarga para el músculo cardiaco, que este intenta compensar al principio mediante la hipertrofia concéntrica del ventrículo izquierdo, caracterizado por un aumento de la pared ventricular. Al final se deteriora la función de esta cámara y la cavidad se dilata, apareciendo los signos y síntomas de la insuficiencia cardiaca. Los hipertensos que desarrollan cardiomegalia en la radiografía de tórax o crecimiento ventricular izquierdo en el electrocardiograma, tienen un riesgo mayor de eventos cardiovasculares.

Estudios animales, anatómicos, epidemiológicos y patogénicos demuestran la relación de la hipertensión arterial con la cardiopatía isquémica. En animales la existencia de hipertensión arterial conlleva aterosclerosis acelerada. Anatómicamente se demuestra que los lechos vasculares perfundidos a altas presiones, como los vasos proximales de la coartación aórtica, tienen mayores lesiones ateroscleróticas. Lo contrario ocurre cuando el vaso es perfundido a baja presión, como es el caso del origen anómalo pulmonar de las arterias coronarias. Los grandes estudios epidemiológicos han demostrado una fuerte asociación entre la hipertensión y la cardiopatía isquémica.

El efecto puede ser la aparición de angina de pecho, a consecuencia de la combinación de enfermedad coronaria acelerada y el aumento de las necesidades miocárdicas de oxígeno, debido al incremento de la masa

miocárdica. La mayor parte de las muertes debidas a la hipertensión son consecuencia de infarto de miocardio o insuficiencia cardiaca congestiva.

1.2.6.2. Efectos sobre el sistema nervioso central.

Dado que la retina es el único tejido en que pueden observarse directamente las arterias, arteriolas y el nervio óptico, constituye una oportunidad para observar los efectos de la hipertensión sobre el árbol vascular. Por ello los efectos neurológicos de la hipertensión arterial pueden dividirse en retinianos y sobre el sistema nervioso central propiamente dicho.

El aumento de la gravedad de la hipertensión se asocia a espasmo focal y estrechamiento general progresivo de las arteriolas retinianas, así como a la aparición de hemorragia, exudado y edema de papila, que darán lugar a escotoma, visión borrosa e incluso ceguera. Estas lesiones y síntomas pueden desarrollarse de forma aguda y revertir con el tratamiento. En cambio las lesiones arterioscleróticas son consecuencia de la proliferación del endotelio y del músculo, y reflejan con precisión las lesiones similares que están produciéndose en otros órganos. Los cambios arterioscleróticos no se desarrollan tan rápidamente como los anteriores ni tampoco regresan de forma apreciable con el tratamiento. Las arterias esclerosadas aparecen, en el examen de fondo de ojo, distorsionadas y comprimen las venas cuando se entrecruzan en la vaina fibrosa común, cambiando la luz que reflejan.

Sobre el sistema nervioso central son comunes síntomas como las cefaleas occipitales, más frecuentes por la mañana. También pueden presentarse: mareo, inestabilidad, tinnitus, alteraciones visuales o síncope.

Pero las manifestaciones más graves se deben a oclusión vascular, hemorragias o encefalopatía.

El infarto cerebral es secundario a la mayor arteriosclerosis observada en los pacientes hipertensos, en tanto que la hemorragia cerebral es consecuencia de la elevación de la presión arterial y del desarrollo de microaneurismas cerebrales. La encefalopatía hipertensiva consiste en el siguiente complejo: hipertensión grave, alteración de la conciencia, aumento de la presión intracraneal, retinopatía con edema de papila y convulsiones, no conociéndose bien su patogenia.

1.2.6.3. Efectos renales.

Las lesiones arterioscleróticas de las arteriolas aferente y eferente y de los ovillos glomerulares son las lesiones vasculares renales más frecuentes en la hipertensión y causan disminución del filtrado glomerular y disfunción tubular. Cuando hay lesiones glomerulares se producen proteinuria y hematuria microscópica, y aproximadamente el diez por ciento de las muertes por hipertensión se deben a insuficiencia renal.

La hipertensión y el envejecimiento actúan de modo sinérgico para producir nefrosclerosis. Disminuye el flujo sanguíneo renal, aumenta la resistencia vascular intrarrenal y se reduce la tasa de filtración glomerular y la capacidad para concentrar la orina. La incidencia de hiperuricemia es elevada en los ancianos con hipertensión esencial no tratada, dado que cuanto menor es el flujo sanguíneo renal, más elevada es la concentración sérica de ácido úrico.

Capítulo II: Métodos Estadísticos.

2.1. Introducción.

En este capítulo se desarrollan varios métodos estadísticos, que serán aplicados a un estudio epidemiológico sobre hipertensión arterial en Lanzarote, contenido en el capítulo tercero. El diseño es de caso-control, siendo la variable de clasificación el indicador de la presencia de hipertensión arterial (HTA). Entre los métodos estadísticos, se utiliza el bootstrap, metodología genérica empleada en varios puntos de esta memoria. El test de la t de Student, el test no paramétrico de Wilcoxon, la familia de transformaciones de Box y Cox, cuya finalidad es conseguir normalidad en un conjunto de datos.

Se introducen las Curvas Receiver Operating Characteristics (ROC), para evaluar el valor discriminante de los marcadores unidimensionales. Cada marcador puede ser una variable observada directamente sobre el individuo, o bien, un *score discriminante* obtenido mediante el análisis discriminante canónico.

Se proponen los modelos de regresión logística con la finalidad de detectar el conjunto de variables que posean valor discriminante. Más concretamente, las odd-ratios ajustadas se obtienen mediante el método *logit*. Usando el bootstrap calculamos factores de corrección en los modelos de regresión logística cuando existe escasez de datos.

En la práctica totalidad de métodos utilizados en esta memoria, se desarrollan implementaciones informáticas específicas para resolver las cuestiones planteadas. En algunos se utilizan paquetes estándar; la computación es la plataforma de apoyo.

2.2. Generalidades sobre los estudios epidemiológicos

Antes de comenzar a exponer el diseño propiamente dicho, se comentan distintos conceptos de Epidemiología, ya que esta ciencia está íntimamente relacionada con esta memoria.

Frost (1927) define Epidemiología como "la ciencia de las enfermedades infecciosas, en tanto que son fenómenos de masas o de grupo, consagrada al estudio de su historia natural y su propagación en el marco de una cierta filosofía". El autor la define como una ciencia inductiva de grupo, concediendo gran peso a su evolución. Evidentemente el concepto ha transcendido de la enfermedad infecciosa, siendo el marco actual más amplio, con lo que la definición queda obsoleta.

Mac Mahon y Pugh (1970) afirman que es "el estudio de la distribución de la enfermedad en el hombre y de los factores que determinan su frecuencia". Se trata de una definición demasiado genérica, si bien encierra gran contenido.

Lilienfeld (1983), "la epidemiología estudia los patrones de distribución de las enfermedades en las poblaciones humanas, así como los factores que influyen en dichos patrones". Se trata una definición en la misma línea de la anterior, algo más concreta, aunque sigue presentando carencias.

Last (1987) la define como "el estudio de la distribución y los determinantes de los estados, o acontecimientos relacionados con la salud de las poblaciones". Sigue siendo demasiado genérica. Se aproxima más al actual concepto.

Jenicek (1993) expone el concepto de esta ciencia del siguiente modo: "Un razonamiento y un método propio de trabajo objetivo en medicina y otras ciencias de la salud aplicadas a la descripción de los fenómenos de la salud, a la aplicación de su etiología y a la búsqueda de métodos de intervención más eficaces". Aparece explícito el concepto de etiología, es más amplia y concreta que las anteriores.

La evolución histórica, basada en la experiencia, tiende a integrar todos los aspectos de esta ciencia. Así surgen definiciones en extenso, como la dada por

Kleinbaum, que considera que la epidemiología:" describe el estado de salud de la población, identifica la magnitud del problema, la frecuencia de ocurrencia entre diferentes grupos y la tendencia de la enfermedad; explica la etiología de las enfermedades; determina los factores de riesgo asociados y los modos de transmisión; predice la magnitud y distribución de la enfermedad en las poblaciones, y controla la enfermedad por medio de medidas preventivas y de erradicación".

Existe una corriente que considera que se trata del estudio de la salud del individuo en relación con su medio, tomando en consideración los aspectos ecológicos que condicionan los fenómenos de la salud y enfermedad en los grupos humanos.

La epidemiología, en su acepción más amplia, trasciende de su origen, las enfermedades infectocontagiosas, para desarrollar una metodología orientada a la investigación de todos los problemas de salud y enfermedad que afecten a las poblaciones. Los estudios epidemiológicos, en cuanto a su temporalidad se refiere, se clasifican en sincrónicos o transversales y diacrónicos o longitudinales. En los estudios transversales se elige una muestra aleatoria de individuos determinándose sobre ellos un conjunto de variables; una de ellas es la indicatriz de la presencia o ausencia de enfermedad, y otras relacionadas con sus potenciales factores de riego. La toma de datos se efectúa en el momento o en un corto espacio de tiempo. Los otros tipos de estudios son los diacrónicos o diferidos en el tiempo. Entre ellos están los longitudinales o de cohortes; esta no es más que un grupo de individuos, con alguna característica común, cuya evolución es observada según transcurre el tiempo (prospectivos), o bien se observa su evolución hacia atrás en el tiempo (retrospectivos) . En el capitulo I de esta memoria se ha mencionado el estudio Framingham, que es un ejemplo clásico de estudio de cohortes prospectivo. Una técnica muy usada, para estudio de relación factor de riesgo-enfermedad, consiste en seleccionar muestras en cada cohorte, que se basan en la exposición o no a hipotéticos factores de riesgo, y diferido en el tiempo observar si contraen o no la enfermedad.

2.3. Estudios de caso-control

Los estudios retrospectivos o de caso-control son estudios longitudinales o diacrónicos en los que una población se clasifica en individuos que padecen una cierta enfermedad (casos) y en individuos sanos (controles). Los datos de estos estudios se obtienen muestreando en cada una de las clases consideradas; a saber: casos y controles. A partir de aquí se investiga retrospectivamente su exposición a los distintos factores de riesgo.

Debemos observar que en este tipo de estudios, los muestreos se llevan a efecto en las poblaciones de casos y controles respectivamente, por lo que los datos obtenidos no tienen en general valor predictivo para la patología considerada (véase 2.3.1.). Sin embargo, cuando se considera como variable dependiente un factor que eventualmente podría asociarse a la enfermedad, el correspondiente modelo permite estimar el valor predictivo para ese factor. De esta cuestión se aportan diversos análisis en el tercer capítulo. En cualquier caso, a través de este tipo de estudios puede estimarse la asociación entre dos factores a través de las correspondientes odd-ratios, ajustando por posibles variables de confusión (véase 2.4).

Los estudios de caso-control tienen la ventaja de poder estudiar varios factores de riesgo simultáneamente, cuestión que también es posible mediante el uso de modelos multidimensionales. Son más cortos en el tiempo que los de cohortes, por lo tanto más asequibles. Su desventaja radica en que no permiten calcular la prevalencia, se prestan más a errores sistemáticos y a mayor confusión entre factores que los de cohortes. Tal cosa sucede porque la información se recoge de forma retrospectiva, pues ha tenido lugar en el pasado, lo que induce a que las fuentes documentales no sean exactamente las mismas; o si se realiza con entrevistas personales, el recuerdo puede no ser percibido de la misma manera. Cuando los casos informan sobre la exposición con distinta validez que los controles, se habla de sesgo de recuerdo.

En un contexto de selección aleatoria, representamos por D el suceso de padecer la enfermedad, y por F el suceso de encontrarse expuesto a un hipotético factor de riesgo. En este tipo de estudios son estimables $\mathbf{P}(F/D)$ y $\mathbf{P}(F/D^c)$ siendo F el suceso que representa la exposición al factor analizado y D la ocurrencia de la

enfermedad en estudio. Si se conoce la prevalencia de la enfermedad $P(D)$, el Teorema de Bayes permite estimar probabilidades del tipo:

$$(2.3.1) \qquad \mathbf{P}(D/F) = \frac{\mathbf{P}(F/D)\cdot\mathbf{P}(D)}{\mathbf{P}(F/D)\cdot\mathbf{P}(D)+\mathbf{P}(F/D^c)\cdot\mathbf{P}(D^c)}$$

Hemos de dejar clara constancia de que en un estudio de este tipo no son evaluables directamente probabilidades del tipo $\mathbf{P}(D/F)$ o $\mathbf{P}(D/F^c)$; solamente es posible utilizando la expresión (2.3.1).

Un estudio cuya finalidad es investigar las causas de una enfermedad, supone tratar variables numéricas y categóricas entre los dos grupos de estudio. Para los primeros se utiliza el *t*-test, medias ajustadas, test de las permutaciones bootstrap, transformaciones de Box y Cox conjunta para caso-control, que se encuentran expuestos en el epígrafe 2.6.1 y siguientes. Para los factores de riesgo categóricos se usa la odd-ratio; cruda en un análisis preliminar y ajustada al utilizar metodología multidimensional, que se expone en el epígrafe 2.6.

2.4. Confusión.

El fenómeno de la confusión aparece cuando la supuesta asociación entre el factor de riesgo y la enfermedad se debe en parte o totalmente a una tercera variable que está alterando los resultados. O bien cuando una asociación real queda oculta por una tercera variable, que es el factor de confusión. En muchos estudios epidemiológicos, la edad se comporta como un factor de confusión, pues muchas de las patologías se encuentran ligadas, entre otros, al citado factor. Consideremos un estudio de caso-control donde las edades en los dos grupos no son homogéneas, estando la edad del grupo de los casos sesgada hacia la derecha; con este planteamiento algunas de las asociaciones encontradas entre la variable de clasificación y otra variable del estudio podría atribuirse a la edad. En el contexto de esta memoria, la HTA y la

hipercolesterolemia podrían explicarse por el hecho de que los hipertensos tienen mayor edad que los normotensos y por este motivo tienen mayor nivel de colesterol.

Existen medidas de asociación, como la odd-ratio, que nos permiten determinar si existe o no una relación, pero debemos cerciorarnos si realmente hay una relación de causa-efecto, o bien se trata de un efecto de confusión. La odd-ratio definida en el epígrafe 2.6.3 es la llamada odd-ratio cruda, que mide la asociación entre F factor de riesgo y D enfermedad; obtenido un valor del mismo significativamente distinto de la unidad, quiere decir que existe relación entre el factor y la enfermedad, pero no tiene en cuenta posibles factores de confusión.

Parece propio, al estimar la medida de asociación entre expuestos al factor y enfermedad, ajustar su valor por el resto de variables consideradas; de este modo eliminamos los posibles efectos de confusión. La odd-ratio así estimada se llama ajustada, y, a diferencia del crudo, tiene en cuanta los factores de confusión. Lo que realmente se hace es ajustar la posible relación, por la supuesta variable de confusión (véase 2.6.4).

Estas ideas sobre confusión llevan, inequívocamente, a la necesidad de considerar todos los factores de la enfermedad dentro de un mismo estudio y considerar sus posibles interrelaciones; solamente de este modo se harán patentes los verdaderos factores de riesgo. Es aquí donde interviene, respondiendo a esta necesidad, el análisis multidimensional desarrollado en el epígrafe 2.9.

2.5. El bootstrap

Una herramienta importante para la obtención de la distribución de probabilidad de estadísticos de interés es el bootstrap. Esta metodología se usará en diferentes puntos de esta memoria. La idea del bootstrap, método autosuficiente, es aproximar la distribución de estadísticos de interés, utilizando exclusivamente los datos. Para ilustrar esta idea, consideremos una distribución de probabilidad $F(x)$ sobre $\mathbf{R}$ con varianza

finita σ^2, y supongamos que se requiere hacer inferencias acerca de su media $\mu = \int_{-\infty}^{\infty} x \cdot F(dx)$. En orden a estimar el parámetro μ se requiere tomar una muestra aleatoria $Y_1, \cdots, Y_n$ de la distribución probabilística $F(x)$. Es bien sabido que el estimador $\hat{\mu}_n = (1/n)\sum_{i=1}^{n} Y_i$ tiene excelentes propiedades como las de ser *centrado, consistente* y con distribución de probabilidad *asintóticamente distribución normal*. Ello permite hacer la afirmación:

$$T_n = \sqrt{n}\frac{\hat{\mu}_n - \mu}{\sigma} \approx N(0,1)$$

Nótese que lo que se afirma es sólo que T_n es *aproximadamente* normal. El error preciso de aproximación viene dado por un desarrollo de Edgeworth de la forma:

(2.5.1) $$\mathbf{P}(T_n \leq y) = \Phi(y) + O(n^{-1/2}) \quad , \text{ uniformemente en } y.$$

siendo $\Phi(y)$ la función de distribución de normal. (véase apéndice I, para cuestiones sobre infinitésimos)

Para pequeñas muestras, el uso de esta aproximación, en contrastes de hipótesis, no permitiría garantizar el error *alpha* y, en intervalos de confianza, las probabilidades de cobertura.

En orden a calcular de una manera mucho más precisa la distribución de probabilidad del estadístico T_n, supóngase que se conoce exactamente la distribución $F(y)$. Mediante métodos de simulación de variables aleatorias, podrían extraerse de $F(y)$ todas las muestras de tamaño n que se deseen. Supóngase que se extraen B muestras de tamaño n cada una. Para cada una de estas muestras, se tiene una observación de la variable aleatoria T_n. Los B valores de T_n dan una idea muy clara de esta distribución de probabilidad. Ahora bien, el problema surge del desconocimiento de $F(y)$. Ni siguiera podemos hacer uso de la normalidad, dado que la exploración de los

datos desaconseja hacer uso de esta hipótesis. Sin embargo, una vez obtenida la muestra aleatoria $Y_1,\cdots,Y_n$, puede estimarse $F(y)$, por ejemplo, mediante la distribución empírica, la cual se define por:

$$\hat{F}_n(t)=\frac{1}{n}\sum_{i=1}^{n} I(Y_i \le t)$$

Aquí, $I(A)$ es una variable aleatoria que toma el valor 1 cuando ocurre A y cero en caso contrario. El estimador $\hat{F}_n(t)$ es centrado y consistente para $F(y)$. Estamos entonces en condiciones de extraer B muestras aleatorias $Y_1^*,\cdots,Y_n^*$ de $\hat{F}_n(t)$. A partir de ahí, puede obtenerse la aproximación buscada a la distribución del estadístico T_n. Téngase en cuenta lo siguiente:

1. $E_*\left[Y_i^*\right]=\int_{-\infty}^{\infty} y\hat{F}_n(dy)=\frac{1}{n}\sum_{i=1}^{n} Y_i=\hat{\mu}_n$
2. $\mathrm{var}\left(Y_i^*\right)=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{\mu}_n\right)^2$.

La idea del bootstrap es aproximar el estadístico T_n por su *análogo*:

$$T_n^*=\sqrt{n}\,\frac{\bar{Y}^*-\hat{\mu}_n}{S}$$

siendo $\bar{Y}^*=(1/n)\sum_{i=1}^{n} Y_i^*$ y $S^2=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{\mu}_n\right)^2$. Las B muestras aleatorias $Y_1^*,\cdots,Y_n^*$ de la distribución empírica $\hat{F}_n(t)$ conducen a B valores T_n^* que proporcionan la aproximación buscada a la distribución probabilística de T_n.

Cabe preguntase por el valor de B, número de generaciones bootstrap a realizar. Esta es una de las características de este método: el número de generaciones no es muy grande, y a partir de un cierto número, nuevas generaciones aportan muy poco o nada. Cuando $B \to \infty$, se dice que el bootstrap es ideal. No obstante, dada la alta velocidad de computación que alcanzan los ordenadores de última generación, y el bajo coste de la

misma, se recomienda como aproximación al bootstrap ideal tomar $B=10*n$, siendo n el tamaño de la muestra original. Para mayores detalles ver P. Hall (1986).

El algoritmo se puede expresar en la forma:

Paso 1. Seleccionar una muestra aleatoria $Y_1,\cdots,Y_n$ de la distribución de probabilidad $F(x)$ y elegir una distribución de remuestreo (por ejemplo, la distribución empírica $\hat{F}_n(t)$).

Paso 2. Extraer B muestras aleatorias bootstrap $Y_1^*,\cdots,Y_n^*$ de la distribución de remuestreo considerada.

Paso 3. En cada muestra, obtener un valor del pivotal bootstrap definido por:

$$T_n^* = \sqrt{n}\frac{\overline{Y}^* - \hat{\mu}_n}{S}$$

con lo que tendremos B valores, los cuales le suministrarán una aproximación a la distribución de probabilidad del estadístico T_n.

Bajo condiciones específicas, puede probarse el siguiente desarrollo de Edgeworth (véase P. Hall, 1992, página 84):

$$\mathbf{P}\left(T_n^* \le y \mid Y_1,\cdots,Y_n\right) = \mathbf{P}\left(T_n \le y\right) + O_P\left(n^{-1}\right) \tag{2.5.2}$$

uniformemente en x. (Véase Apéndice I, para cuestiones sobre infinitésimos)

Nótese que la verdadera función de distribución de probabilidad del estadístico T_n, $\mathbf{P}\left(T_n \le y\right)$ se aproxima por la función de distribución de probabilidad bootstrap, $\mathbf{P}\left(T_n^* \le y \mid Y_1,\cdots,Y_n\right)$ con un error del orden $O_P\left(n^{-1}\right)$, lo que significa una mejor aproximación.

Obsérvese que las aproximaciones bootstrap pueden mejorar las aproximaciones a la distribución normal que proporciona el teorema central del límite. A este respecto consultar 2.5.1 y 2.5.2. Aunque no es objeto de esta memoria, las aproximaciones bootstrap resultan aún de mayor utilidad en aquellos casos en los que no es aplicable el teorema central del límite. Ello ocurre, por ejemplo, cuando se aproxima la distribución de probabilidad de estimadores no lineales.

2.6. Análisis de potenciales factores de riesgo.

Comenzamos realizando un análisis preliminar o previo, que considera los potenciales factores de riesgo de modo individual, sin tener en cuenta sus posibles interacciones. Proporciona indicios sobre posibles asociaciones con la enfermedad, diferencia de medias entre casos y controles, así como otros estadísticos de interés. Este tipo de análisis no es completo, pues no considerando todos los factores en su conjunto, cuestión de la que nos ocuparemos más tarde en el análisis multivariante. Se realiza en dos vertientes, la primera considera los factores numéricos y la segunda considera solo los factores categóricos.

2.6.1. Análisis de potenciales factores de riesgo numéricos.

Se consideran en este epígrafe solamente los potenciales factores de riesgo de carácter numérico. En este análisis preliminar se emplean varias metodologías para comparar medias entre los dos grupos: *t*-test, test de las permutaciones bootstrap, pruebas no paramétricas y transformaciones de Box y Cox; cuestión esta última, desarrollada ampliamente en esta memoria.

2.6.1.1. Comparación de medias mediante el t-test.

En aquellos supuestos en que los datos se distribuyan normalmente en ambos grupos de estudio, éstos quedan bien resumidos por la media y desviación estándar, y por tanto, la comparación de medias es el modo más adecuado de comparar las variables numéricas de interés. El test de la *t* de Student o *t*-test es, en este supuesto, el método

más eficiente para la comparación de medias, garantizando el error de tipo *alpha* admitido por el investigador.

Sean por tanto μ_D y μ_C las medias de la característica numérica de interés en las poblaciones de enfermos (casos) y controles respectivamente. Representaremos por $Y_{i,1},\cdots,Y_{i,n_i}$ los datos observados en el grupo *i*-ésimo (*i=D,C*). La hipótesis de normalidad supone que $Y_{i,j} \cong N(\mu_i,\sigma_i)$ (los datos son generados por una ley de probabilidad con media μ_i y desviación estándar σ_i). La comparación *cruda* de la característica numérica entre ambas poblaciones supone (en el supuesto de normalidad) comparar las medias μ_D y μ_C. La hipótesis nula a contrastar tiene la forma $H_0 : \mu_D = \mu_C$. La hipótesis alternativa puede ser de tipo bilateral ($H_1 : \mu_D \neq \mu_C$) o de tipo unilateral ($H_1 : \mu_D > \mu_C$). En aquellos casos en que pueda admitirse la hipótesis de homoscedasticidad ($\sigma_D = \sigma_C = \sigma$), es fácil probar que, supuesto que la hipótesis nula H_0 sea cierta, el test estadístico:

$$Z = \frac{\bar{Y}_D - \bar{Y}_C}{\sigma\sqrt{1/n_D + 1/n_C}} \cong N(0,1)$$

siendo $\bar{Y}_i = (1/n_i)\sum_{j=1}^{n_i} Y_{i,j}$, *i=D,C*. Este estadístico es poco práctico cuando no se conoce la desviación estándar común σ y procede por tanto el uso del estadístico estudentizado alternativo:

$$T = \frac{\bar{Y}_D - \bar{Y}_C}{S_p\sqrt{1/n_D + 1/n_C}}$$

siendo ahora $S_p^2 = \left(1/(n_D + n_C - 2)\right)\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(Y_{i,j} - \bar{Y}_i\right)^2$, siendo este último estadístico un estimador centrado y de mínima varianza para σ^2. Bajo la hipótesis nula $H_0 : \mu_D = \mu_C$, $T \cong t(n_D + n_C - 2)$ ($t(n)$ representa la ley de probabilidad *t* de Student con *n* grados de libertad). Raramente pueden darse por conocidas las varianzas σ_D^2 y

σ_C^2, por lo que generalmente, el test utilizado para la comparación de medias se basará más en el estadístico T, que en el Z.

El uso de estos estadísticos supone la hipótesis de homoscedasticidad que previamente debe ser contrastada. Por tanto, la comparación de medias debe estar precedida de una comparación de varianzas de la forma $H_0 : \sigma_D = \sigma_C$ frente a la alternativa $H_1 : \sigma_D \neq \sigma_C$. Para esta comparación procede el uso del test estadístico de Levene, que tiene la forma:

$$F = \frac{\frac{1}{n_D - 1}\sum_{j=1}^{n_D}\left(Y_{D,j} - \bar{Y}_D\right)^2}{\frac{1}{n_C - 1}\sum_{j=1}^{n_C}\left(Y_{C,j} - \bar{Y}_C\right)^2}$$

el cual, bajo la hipótesis nula $H_0 : \sigma_D = \sigma_C$, sigue una ley de probabilidad $\mathsf{F}\left(n_D - 1; n_C - 1\right)$ (ley de probabilidad F de Snedecor).

Si se admite la hipótesis de homoscedasticidad y suponiendo que las varianzas σ_D^2 y σ_C^2 son desconocidas, para la comparación de medias, una vez observado un valor t_0 del estadístico T, son óptimos los test cuyo p-valor o significación se obtiene como:

a. Para el contraste bilateral: $p = \mathbf{P}_{H_0}\left(\,|T| > |t_0|\,\right)$

b. Para el contraste unilateral: $p = \mathbf{P}_{H_0}\left(\,T > t_0\,\right)$

La notación $\mathbf{P}_{H_0}$ representa la medida de probabilidad bajo la hipótesis nula $H_0 : \mu_D = \mu_0$.

En los supuestos de heteroscedasticidad ($\sigma_D \neq \sigma_C$), el test estadístico T debe sustituirse por el alternativo:

$$T_S = \frac{\bar{Y}_D - \bar{Y}_C}{\sqrt{S_D^2/n_D + S_C^2/n_C}}$$

siendo $S_i^2 = 1/(n_i - 1)\sum_{j=1}^{n_i}(Y_{i,j} - \bar{Y}_i)^2$, $i=D,C$. Bajo la hipótesis nula $T_s \cong t(n_D + n_C - 2)$

En aquellos casos en los que se mantiene la homoscedasticidad (o ésta puede razonablemente admitirse), la potencia de cada uno de los test puede evaluarse en función del parámetro $\theta = (\mu_D - \mu_C)/\sigma$. Para el contraste unilateral, la hipótesis alternativa puede obviamente expresarse en la forma $H_1 : \theta > 0$ y la función de potencia del test de tamaño α en la forma $\mathbf{P}_\theta(T > t_\alpha(n_D + n_C - 2))$ ($t_\alpha(n)$ es el cuantil 1-α de la distribución t de Student con n grados de libertad). Para el contraste bilateral, la forma de la función de potencia del test de tamaño α es $\mathbf{P}_\theta(\,|T| > t_{\alpha/2}(n_D + n_C - 2)\,)$.

La consideración de la función de potencia puede ser un aspecto de interés en el diseño del estudio. Pensamos que cualquier estudio epidemiológico debe tener siempre un objetivo principal o *endpoint* bien definido, sin perjuicio de que se analicen otros factores. En el estudio que se desarrolla en el tercer capítulo de esta memoria, el objetivo principal podría consistir en evaluar la influencia de la obesidad sobre la HTA. Dado que el diseño del estudio es de caso-control, el objetivo se concretaría en probar que en el grupo de enfermos hay una elevación del índice de masa corporal. El investigador podría tener información acerca del parámetro α, y entonces determinar el tamaño de la muestra a seleccionar en cada población en orden a que el test considerado, para un determinado tamaño α, alcance una potencia determinada 1-α para el valor α conjeturado. De la expresión de la función de potencia dada anteriormente, es fácil deducir que los tamaños por grupo requeridos vienen dados por la expresión:

a. Para el contraste bilateral: $n_D = n_C = \dfrac{2(z_{\alpha/2} + z_\beta)^2}{\theta^2}$

b. Para el contraste unilateral: $n_D = n_C = \dfrac{2(z_\alpha + z_\beta)^2}{\theta^2}$

Siendo z_α el cuantil 1-α de la distribución normal estándar.

En el supuesto de obtenerse significación en los contrastes de hipótesis, es obvio que interesará evaluar la magnitud de la diferencia $\delta = \mu_D - \mu_C$. Para tal fin puede construirse un intervalo de confianza para este parámetro basado en la cantidad pivotal:

$$\frac{\bar{Y}_D - \bar{Y}_C - (\mu_D - \mu_C)}{S_p\sqrt{1/n_D + 1/n_C}} \qquad (2.6.1)$$

cuya distribución de probabilidad es $t(n_D + n_C - 2)$ si se mantienen las hipótesis de normalidad y homoscedasticidad. Ello significa que estas hipótesis deben ser previamente contrastadas. Anteriormente se mostró el test de Levene para la comparación de varianzas. Si no puede admitirse la referida hipótesis de homoscedasticidad, se dio un test alternativo para el caso de heteroscedasticidad. La hipótesis de normalidad debe también ser contrastada. El *distanciamiento* de los datos de esta hipótesis podría conducir a que no se pudiera garantizar el error *alpha*. Es frecuente que la normalidad se contraste mediante alguno de los contrastes clásicos de normalidad (Kolmogorov-Smirnov o Saphiro-Wilks). Debe tenerse en cuenta sin embargo, que la hipótesis nula de estos contrastes consiste en afirmar que los datos han sido generados por una ley normal. Si el test utilizado tiene una potencia muy baja debido al escaso tamaño muestral, la conclusión podría ser *no rechazar la hipótesis de normalidad*, aunque esta fuese falsa. Nosotros recomendamos que la hipótesis de normalidad sea explorada gráficamente antes de realizar cualquier contraste de hipótesis. En este sentido, es recomendable el uso de los diagramas de cajas y barras, histogramas y P-plot.

Esta última representación gráfica, P-plot, es del tipo x-y; en el eje horizontal figuran las probabilidades acumuladas observadas, procedentes de los datos experimentales. En el eje vertical se representan las probabilidades acumuladas esperadas en el caso que la distribución fuera normal. Así, para cada valor de la variable obtenemos dos probabilidades: la estimada a partir de los datos y la obtenida si los datos hubiesen sido generados por una ley normal. Evidentemente la distribución normal se representará en esta gráfica mediante una recta, que es la bisectriz del primer

cuadrante, pues probabilidades acumuladas esperadas y observadas coinciden. La normalidad de los datos se mide según los puntos se agrupen entorno a la recta.

La cuestión que hay que plantear en este punto es la siguiente: ¿qué ocurre si los datos se *distancian* de la hipótesis de normalidad ?. A esta cuestión suele dársele dos soluciones alternativas. La primera es *apelar* al efecto del teorema central del límite. Los estadísticos que se han utilizado para los contrastes de hipótesis o para la construcción de intervalos de confianza para $\mu_D - \mu_C$ son lineales en las observaciones y por tanto, susceptibles de que puedan aplicárseles el referido efecto del teorema central del límite. Algunos autores afirman que para tamaños muestrales superiores a 30 la aproximación es aceptable. Otra forma de resolver la cuestión es *optando* por los contrastes de distribución libre, tales como los test de Wilcoxon.

No negamos que la justificación del *t*-test es válida para tamaños muestrales suficientemente amplios, aunque el tamaño 30 posiblemente sea insuficiente. No obstante, pueden obtenerse mejores aproximaciones a la distribución de estadísticos del tipo Z ó T (la versión estudentizada de Z).

El error de la aproximación normal a la distribución de probabilidad del estadístico Z viene dada por los desarrollos de Edgeworth. Para fijar ideas, consideremos el estadístico Z dado por:

$$Z = \frac{\overline{Y}_D - \overline{Y}_C - (\mu_D - \mu_C)}{\sigma\sqrt{1/n_D + 1/n_C}}$$

Ahora suponemos que los datos observados $Y_{i,1},\cdots,Y_{i,n_i}$ siguen una ley de probabilidad no conocida de media μ_i (i=D,C) y desviación estándar común σ. la cual es finita. De acuerdo con los referidos desarrollos de Edgeworth, podemos afirmar:

$$\mathbf{P}(Z \le x) = \Phi(x) + O(n^{-1/2}) \qquad (2.6.2)$$

Ello significa, que para tamaños muestrales del orden de 100, el error de aproximación es del orden de una décima. La metodología bootstrap introducida por Efron en 1979 como alternativa al Jackknife (Quenouille, 1957), B. Efron (1979) y B. Efron and Tibshirani (1986), puede, bajo determinadas condiciones, conducir a mejores aproximaciones. En el siguiente epígrafe haremos una breve revisión de esta metodología.

2.6.1.2. Aproximación bootstrap para la diferencia de medias.

Supongamos ahora que la distribución de la característica de interés tiene distribuciones de probabilidad F_D y F_C en las poblaciones de enfermos y sanos respectivamente y se desea estimar la diferencia de medias $\mu_D - \mu_C$ mediante un intervalo de confianza. Aquí, $\mu_i = \int y \cdot F_i(y)$, $i=D,C$. El intervalo puede construirse a partir de la cantidad pivotal:

$$T = \frac{\overline{Y}_D - \overline{Y}_C - (\mu_D - \mu_C)}{S_p \sqrt{1/n_D + 1/n_C}}$$

La aproximación clásica dada por el teorema central del límite, permite aproximar esta distribución por una $t(n_D + n_C - 2)$. Procedemos, sin embargo, a aproximarla mediante el siguiente algoritmo bootstrap:

Paso 1. De cada distribución de probabilidad F_i, $i=D,C$ seleccionar una muestra aleatoria $Y_{i,1},\cdots,Y_{i,n_i}$ y construir las correspondientes distribuciones empíricas

$\hat{F}_{D,n_D}(t)$ y $\hat{F}_{C,n_C}(t)$. Sea $\overline{Y}_i = (1/n_i)\sum_{j=1}^{n_i} Y_{i,j}$.

Paso 2. Extraer B muestras aleatorias $Y^*_{i,1},\cdots,Y^*_{i,n_i}$ de la distribución de remuestreo $\hat{F}_{i,n_i}(t)$ para $i=D,C$.

Paso 3. Obtener B valores del pivotal bootstrap definido por:

$$(2.6.3.) \qquad T^* = \frac{\bar{Y}_D^* - \bar{Y}_C^* - \left(\bar{Y}_D - \bar{Y}_C\right)}{S^* \sqrt{1/n_D + 1/n_C}}$$

siendo $\bar{Y}_i^* = \left(1/n_i\right)\sum_{j=1}^{n_i} Y_{i,j}^*$. Los B valores de T^* proporcionan la deseada aproximación a la distribución de probabilidad de la cantidad pivotal T.

2.6.1.3. Estudio de simulación.

Para ilustrar las aproximaciones proporcionadas por diferentes métodos expuestos en los epígrafes 2.5. y 2.6.1.1, consideramos una distribución de probabilidad que se aparte fuertemente de la normalidad. Proponemos para el estudio de simulación la distribución exponencial con parámetro λ=0.2, cuya densidad tiene la forma $f(x) = \lambda \cdot \exp(-\lambda x)$; $x \geq 0$. Para una muestra aleatoria de esta distribución $X_1, \cdots, X_n$, analizamos el estadístico $T_n = \sqrt{n} \cdot \left(\bar{X} - \mu\right)/S$.

Tabla 2.6.4. Estudio comparativo de aproximación (Datos simulados).

		$P_{2.5}$	P_{25}	P_{50}	P_{75}	$P_{97,5}$
REAL		-2.59	-0.79	-0.04	0.60	1.69
Aproximación T		-2	-0.68	0	0.68	2
Aproximaciones	1	-2.56	-0.80	-0.05	0.60	1.68
bootstrap	2	-2.43	-0.80	-0.03	0.60	1.69
	3	-2.56	-0.66	0.07	0.72	1.82
	4	-2.56	-0.93	-0.11	-0.58	1.68
	5	-2.79	-0.87	-0.11	-0.65	1.74
	6	-2.74	-0.87	-0.05	-0.54	1.79
	7	-2.41	-0.69	-0.05	0.62	1.76
	8	2.45	-0.77	-0.08	0.59	1.78
	9	-2.47	-0.75	0.01	0.69	1.64
	10	-2.17	-0.64	0.11	0.71	2.03

La Tabla 2.6.4 muestra los percentiles correspondientes a la distribución de T_n, para n=50, así como los correspondientes a la distribución t de Student. Como puede observarse, los percentiles correspondientes a la distribución t se apartan notablemente de los reales. Mostramos ahora como diversas aproximaciones bootstrap mejoran la aproximación t. Para ello tomamos el número de réplicas bootstrap $B = 10 \cdot n$ (500). De las 10 generaciones realizadas, en todas ellas la aproximación bootstrap mejora la obtenida por la t de *Student*.

2.6.1.4. El test de las permutaciones bootstrap.

Se plantea en este epígrafe, otra prueba comparativa de medias, el test de las permutaciones bootstrap. Se trata de un método no paramétrico o de libre distribución, que, al contrario del *t*-test que suponía normalidad, no presupone ninguna forma específica en datos. Así los contrastes de distribución libre se basan en conceptos como orden, rango o signos; y son útiles cuando *la distancia* de los datos a normalidad es grande, o bien como prueba comparativa con otra paramétrica. Es en este último sentido como se emplea en esta memoria, pues los resultados de ambos métodos se presentan conjuntamente.

Representaremos por $Y_{i,1},\cdots,Y_{i,n_i}$ los datos observados en el grupo *i*-ésimo (*i*=*D*,*C*). Si bien los tamaños muestrales n_D y n_C no tienen que coincidir. Contrastamos: $H_0 : \mu_D = \mu_C$ y como hipótesis alternativa $\mu_D \neq \mu_C$, o bien $\mu_D > \mu_c$.

Bajo la hipótesis nula las dos medias tienen la misma distribución $F_D(Y) = F_C(Y) = F(Y)$ donde F es desconocida. El *p*-valor genérico viene dado por la expresión $p = \mathbf{P}_{H_0}\left(T \geq t : S = s\right)$, en nuestro caso es $t = \left| \overline{Y}_{n_D} - \overline{Y}_{n_C} \right|$.

El siguiente algoritmo bootstrap ilustra el desarrollo de esta prueba:

Paso 1. Ordenamos los datos de los dos grupos en un solo conjunto, así $S = \left\{ Y_{D,1},\cdots,Y_{D,n_D},Y_{C,1},\cdots,Y_{C,n_C} \right\}$.

Paso 2. Se seleccionan aleatoriamente n_D datos. Se estima la media bootstrap $\overline{Y}^*_{n_D}$.

Paso 3. Se seleccionan aleatoriamente n_c datos. Se estima la media bootstrap $\overline{Y}^*_{n_c}$.

Paso 4. Comparar el valor absoluto de las medias bootstrat $\left| \overline{Y}^*_{n_c} - \overline{Y}^*_{n_D} \right|$, con la misma diferencia referida a la medias muestrales $\left| \overline{Y}_{n_C} - \overline{Y}_{n_D} \right|$, si la diferencia de medias bootstrap es mayor que la diferencia de medias muestrales, $\left| \overline{Y}^*_{n_c} - \overline{Y}^*_{n_D} \right| > \left| \overline{Y}_{n_C} - \overline{Y}_{n_D} \right|$, decimos que se ha producido una permutación.

Paso 5. Repetir los pasos del 2 al 4 B veces. Considerando en cada una de ellas si se ha producido una permutación.

Paso 6. Se evalúa el estadístico $p = \frac{1 + n^{o}\ permutaciones}{B+1}$; este es el p-valor del test.

Hemos de reiterar que se trata de una prueba empírica, sin supuestos previos, si bien tiene un neto carácter numérico, cuestión que la diferencia de otras pruebas no paramétricas, basadas en otros conceptos, como es el test de Wilcoxon, que se expone a continuación.

2.6.1.5. El test de Wilcoxon para muestras independientes.

Como ya se ha comentado en el epígrafe 2.6.1.1, esta prueba complementa al *t*-test. Esta alternativa de libre distribución, está basada en el concepto de orden y no solo se aplica a datos numéricos, sino a cualquier otro tipo susceptible de ordenación, siendo también conocida como la prueba de los rangos de Wilcoxon. No vamos a contrastar la igualdad de medias, como se hacía en el test paramétrico o en el de las permutaciones boostrap, simplemente se contrasta el hecho de encontrarse ante dos poblaciones diferentes.

Representaremos por $Y_{i,1}, \cdots, Y_{i,n_i}$ los datos observados en el grupo *i*-ésimo ($i=D,C$). Si bien los tamaños muestrales n_D y n_C no tienen que coincidir. Contrastamos:

H_0 : "Los datos provienen de la misma población"

H_1 : "Los datos provienen de distintas poblaciones"

Planteado así, se trata de un contrate de dos colas, si bien admite alternativas en la hipótesis de trabajo, en cuanto a si una población está situada por encima de la otra. La definición del test estadístico es la siguiente:

Paso 1. Los $n_D + n_C$ datos se reúnen en una única muestra ordenada. A cada elemento le asignamos un número de orden, al que llamamos *rango*. Si existen valores repetidos, empates, les asignamos un rango promedio. Cada valor, a pesar de estar en una muestra común no pierde la identificación de su grupo de procedencia.

Paso 2. Si $n_D < n_C$, sin pérdida de generalidad, calculamos el estadístico $W_{n_D} = \sum_{i=1}^{n_D+n_C} R_{i,n_D}$, donde R_{i,n_D} es el *i*-ésimo rango del grupo *D*, con $i = 1,\cdots,n_D$. Si es $n_D > n_C$, se realiza el procedimiento análogo con W_{n_C}.

Paso 3. Evidentemente, si estamos ante datos procedentes de una misma población los valores de W_{n_D} (o en su caso de W_{n_C}) serán moderados. Mientras que valores pequeños del estadístico, indicarían que la población *D*, en la característica evaluada, se encuentra por debajo de la población *C*. Si el estadístico toma valores grandes, se tiene lo contrario. La concreción numérica de valores altos o pequeños se encuentran tabulados en intervalos de aceptación hasta $n_D + n_C \leq 30$. Si se sobrepasa este valor, la variable aleatoria *W*=suma de los rangos de menor tamaño (bien sea W_{n_D} o W_{n_C}), bajo la hipótesis nula, está normalmente distribuida con esperanza y varianza dadas por:

$$E[W] = \left((n_D + n_C + 1)/2\right) \cdot n_D$$

La varianza toma valores diferentes según existan o no empates. En este último caso su valor es:

$$\text{var}(W) = \left((n_D + n_C + 1)/12\right) \cdot n_D \cdot n_C$$

En caso de existir empates, toma la forma:

$$\text{var}(W) = \frac{(n_D + n_C)\left\{(n_D + n_C)^2 - 1\right\} \sum_i T_i}{12(n_D + n_C)} \cdot \frac{n_D \cdot n_C}{n_D + n_C - 1}$$

donde el término $\sum_i T_i$ representa la sumatoria de los empates, con $T_i = t_i^3 - t_i \quad \forall i$ empate; t_i representa el número de datos iguales en el *i*-ésimo empate.

Si $Y_{i,1},\cdots,Y_{i,n_i}$ $(i = D,C)$, siguen una distribución normal, entonces este test contrasta exactamente la misma hipótesis que el *t*-test.

Existen otros contrates de libre distribución que son equivalentes al test de Wilcoxon. El más conocido es el test de Mann-Withney, que también depende de las observaciones linealmente ordenadas. Se evalúa el estadístico U = *número de veces que un valor* Y_D, *precede a un valor* Y_C; así, si la población Y_D, está por encima de la Y_C, el valor de U será grande. Si la situación es la inversa U será pequeño; valores moderados indican que no existe diferencia entre las poblaciones.

2.6.1.6. Transformaciones de Box y Cox.

Una cuestión fundamental en el análisis de datos biomédicos es lograr una forma adecuada de resumir una variable numérica. Está plenamente aceptado que cuando los datos obedecen a una ley normal, la media junto a la desviación estándar son los parámetros que mejor resumen la distribución. Lamentablemente, es frecuente que las características numéricas se *distancien* notablemente de la normalidad. Por tal motivo, es de gran interés transformar los datos de tal forma, que una vez obtenidos los transformados, estos puedan considerarse normalmente distribuidos.

Así pues, considérese una variable aleatoria X y sea la transformación monótona creciente tal que $Y = g(X)$ tal que $Y \approx N(\mu, \sigma)$. La normalidad de esta variable nos permite hacer directamente afirmaciones del tipo:

$$\mathbf{P}\left(\mu - z_{\alpha/2}\sigma \leq Y \leq \mu + z_{\alpha/2}\sigma\right) = 1 - \alpha$$

Esto da ideas muy precisas y prácticas acerca del rango de la variable. Naturalmente, el usuario demanda expresar la variable en sus unidades originales. Dado que la transformación es monótona, existirá su función inversa g^{-1}. De esta forma, podemos escribir:

$$\mathbf{P}\left(g^{-1}\left(\mu - \sigma z_{\alpha/2}\right) \leq X \leq g^{-1}\left(\mu + \sigma z_{\alpha/2}\right)\right) = 1 - \alpha$$

De esta forma, parece claro que un modo adecuado de resumir la variable es mediante el parámetro $g^{-1}(\mu)$.

Una de las familias de transformaciones frecuentemente utilizadas es la debida a Box y Cox (ByC, 1964), que tiene la forma:

(2.6.5.)
$$y = g_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & : \quad \lambda \neq 0 \\ \log x & : \quad \lambda = 0 \end{cases}$$

donde el parámetro λ es la llamada *potencia* de ByC. Si la variable original X es positiva y tiene distribución de probabilidad absolutamente continua con función de densidad de probabilidad dada por $f_X(x)$, la variable transformada $Y = g(X)$ es también absolutamente continua con función de densidad de probabilidad $f_Y(x)$, siendo $f_X(x) = f_Y\left(g_\lambda(x)\right) \cdot \left|g'_\lambda(x)\right|$, lo que supone para $\lambda \neq 0$:

$$f_X(x) = f_Y\left(\frac{x^\lambda - 1}{\lambda}\right) \cdot \left|x^{\lambda-1}\right|$$

En el estudio de caso-control se requiere tomar la misma transformación para ambos subgrupos. Más concretamente, sea $\left\{x_{i,j} \ : i = D,C;\ j = 1,\cdots,n_i\right\}$ el conjunto de datos disponibles para el estudio y sea la transformación $y_{i,j} = \left(x_{i,j}^\lambda - 1\right)/\lambda$, suponiendo que λ es tal que $y_{i,j} \cong N(\mu_i, \sigma)$. La log-verosimilitud correspondiente a los datos transformados será entonces:

(2.6.6.)
$$l\left(\lambda, \mu_D, \mu_C, \sigma^2\right) = -\frac{n_D + n_C}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=D,C} \sum_{j=1}^{n_i} \left(y_{i,j} - \mu_i\right)^2 + (\lambda - 1) \sum_{i=D,C} \sum_{j=1}^{n_i} x_{i,j}$$

Maximizando la verosimilitud para λ fijo se obtiene:

$$\hat{\mu}_i(\lambda) = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} y_{i,j} \quad \text{y} \quad \hat{\sigma}^2(\lambda) = \frac{1}{n_D + n_C} \cdot \sum_{i=D,C} \sum_{j=1}^{n_i} \left(y_{i,j} - \mu_i\right)^2$$

Sustituyendo $\hat{\mu}_i(\lambda)$ y $\hat{\sigma}^2(\lambda)$ en la expresión de la log-verosimilitud se obtiene:

(2.6.7.)

$$l(\lambda) = -\frac{n_D + n_C}{2}\log\left(2\pi\hat{\sigma}^2(\lambda)\right) - \frac{1}{2\hat{\sigma}^2(\lambda)}\sum_{i=D,C}\sum_{j=1}^{n_i}\left(y_{i,j} - \hat{\mu}_i(\lambda)\right)^2 + (\lambda - 1)\sum_{i=D,C}\sum_{j=1}^{n_i} x_{i,j}$$

La potencia de Box-Cox estimada $\hat{\lambda}$ es entonces aquella que maximiza la log-verosimilitud $l(\lambda)$. En la práctica, se sabe que $\hat{\lambda}$ no toma valores muy altos, pues estos deformarían totalmente los datos y no se conseguiría normalidad; aún más generalmente los valores se encuentran próximos a cero. Por esta razón, la variación de $\hat{\lambda}$ solamente se realiza en un intervalo pequeño que contenga a cero. Señalar, orientativamente [-4, 4].

Se tiene además:

$$2\left(l(\hat{\lambda}) - l(\lambda)\right) \approx \chi^2(1)$$

Si fijamos un nivel de confianza 1- α, se obtiene una acotación para la diferencia de las verosimilitudes:

$$\left(l_{máx}(\lambda_0) - l(\lambda)\right) \leq \frac{\chi_\alpha{}^2(1)}{2}$$

Lo que equivale a:

(2.6.8.) $$l(\lambda) \geq l_{máx}(\lambda_0) - \frac{1}{2}\chi_\alpha^2(1)$$

Por lo tanto para un valor de la función $l_{máx}(\lambda_0) - \frac{1}{2}\chi_\alpha^2(\alpha)$, existirán dos valores el parámetro λ_1 y λ_2, que formarán un intervalo de confianza (λ_1, λ_2) a un nivel de confianza 1-α, para *la potencia*. Si el valor $\lambda_0 = 1$, está incluido en el intervalo de confianza, la transformación podría ser una traslación, que no afecta a la naturaleza de los datos.

2.6.1.7. Discusión.

Una vez expuestas las diversas posibilidades de análisis de factores numéricos, epígrafe 2.6.1, la pregunta se plantea de un modo natural: ¿qué método elegir, cual proporciona mejores resultados?.La respuesta no es obvia, se elegirá uno u otro dependiendo de cómo sean los elementos a tratar.

Si los datos siguen aproximadamente una distribución normal, lo más apropiado es el *t*-test; aún más, si los datos son escasos este procedimiento proporciona mejores resultados. El supuesto de normalidad deberá ser siempre comprobado mediante alguna prueba; se recomienda ser especialmente cuidadoso en este aspecto, pues un número muy escaso de datos confiere a los test poca potencia y hace poco fiable la opción por la hipótesis nula (de normalidad en nuestro caso). Aún aceptando la hipótesis de normalidad, si tenemos un gran volumen de datos los resultados proporcionados por el bootstrap son prácticamente equivalentes al *t*-test.

Si los datos se *distancian* de la hipótesis de normalidad, parece recomendable intentar una transformación que los acerque a esta hipótesis. De conseguirlo se realizará el tratamiento de los datos transformados a normales, y una vez obtenidos los resultados se llevara a cabo la transformación inversa. En este caso también el bootstrap proporciona buenos resultados.

En el caso de que la *distancia* a normalidad sea muy grande, lo más recomendable es emplear una herramienta no paramétrica, como puede ser el test de Wilcoxon.

Es de destacar la bondad de la metodología bootstrap, pues en la mayoría de supuestos proporciona estimaciones muy similares a las del método en principio recomendado; cuestión esta computacionalmente demostrada.

Todo lo dicho en este epígrafe sobre elección de la prueba, es aplicable no solo a la bondad de los resultados del test en cuanto a significación, sino también es extensible a los intervalos de cobertura y al error cometido.

2.6.2. Medias ajustadas

El hecho de que la media de la característica numérica evaluada en el grupo de casos sea significativamente diferente a la del grupo control, no significa que exista una relación de causalidad entre la característica considerada y la variable binaria de clasificación. El problema de la confusión ya se ha expuesto en el epígrafe 2.4. Supóngase por ejemplo que se está analizando si el nivel medio de la LDL es mayor en el grupo de los hipertensos que en el de los normotensos. Una explicación ingenua o precipitada podría ser que el aumento en el nivel de la LDL implica, debido a un sencillo mecanismo fisiopatológico, una elevación de la tensión arterial. Por otra parte, parece plausible que la edad media de los hipertensos sea superior a la de los normotensos. Dado que los niveles lipídicos tienen a crecer con la edad, la asociación encontrada en una primera instancia podría atribuirse total o parcialmente al factor de confusión *edad*. Por ello, podría ocurrir que en el grupo de hipertensos el mayor nivel medio de la LDL sea atribuible solamente a la mayor edad. En este caso, las medias no son realmente comparables debido a la heterogeneidad causada por las diferencias de edades. Para que estas medias fuesen realmente comparables, habría que hacerlo dentro de cada una de las edades. Esto puede hacerse mediante el siguiente modelo de análisis de la covarianza.

Supóngase ahora que se dispone del siguiente conjunto de datos:

$$\left\{\left(X_{i,j};Y_{i,j}\right) : i = D,C \; ; \; j = 1,\cdots,n_i\right\}$$

El objetivo es comparar la media de la distribución de las variables $Y_{D,j}$ con la de las variables $Y_{C,j}$, para el mismo valor de la supuesta variable de confusión . En el ejemplo anterior, las variables Y representan niveles de la LDL y las X, las edades. Para el fin propuesto, considérese el modelo de análisis de la covarianza:

$$Y_{i,j} \cong N\left(\theta + \alpha_i + \beta \cdot X_{i,j}, \sigma\right) \; : \; i = D,C \; ; \; j = 1,\cdots,n_i$$

Habitualmente, $\alpha_D = \alpha$ y $\alpha_C = 0$

2.6.3. Análisis de potenciales factores de riesgo categóricos: *odd-ratio.*

Se trata de aplicar pruebas de relación y estimar sus *odd-ratios crudas*. De este modo se obtendrá una primera idea de los potenciales factores de riesgo y su influencia.

Las medidas de asociación en general, y sobre todo el riesgo relativo, tienen una fácil interpretación; pero no siempre son estimables. Por ello debemos acudir a otra medida, no tan sencilla de interpretar, pero cuya estimación puede ser realizada en todo tipo de estudios; se trata de la odd-ratio (razón de los impares) o razón del producto cruzado de Yule y también conocida como razón de las ventajas. La definición, en términos probabilísticos, es la siguiente:

$$OR = \frac{\mathbf{P}\left(D|F\right)\cdot\mathbf{P}\left(D^c|F^c\right)}{\mathbf{P}\left(D^c|F\right)\cdot\mathbf{P}\left(D|F^c\right)} = \frac{\mathbf{P}\left(F|D\right)\cdot\mathbf{P}\left(F^c|D^c\right)}{\mathbf{P}\left(F^c|D\right)\cdot\mathbf{P}\left(F|D^c\right)}$$

donde F es el factor, e indica los expuestos al mismo, y D la enfermedad.

Cuando OR=1, no existe asociación entre la enfermedad de estudio y el factor examinado. Esta asociación es positiva cuando OR>1 y negativa o inversa para OR<1. Como puede observarse en la definición anterior, las probabilidades que aparecen en el primer cociente son estimables en un estudio de prospectivo de cohortes, mientras que las que aparecen en el segundo, lo son en estudios de caso-control. Ambos tipos de probabilidades son también estimables en estudios transversales. De esta forma, la odd-ratio, al contrario que el riesgo relativo, es una medida de asociación estimable en cualquier tipo de estudio.

Expresamos como F el hipotético factor de riesgo y D el suceso padecer la enfermedad. El tamaño de la muestra es n; donde n_1, n_2, n_D, n_{D^c}, son los totales para expuestos o no expuestos al factor, enfermos y sanos, respectivamente. Conocemos, como información básica, el número de enfermos, expresadas en la tabla en columnas. Por lo tanto, para cualquier celda de la tabla, pongamos por caso n_{11}, se tiene que: $n_{11} \cong b\left(n_D, \pi_D\right)$; donde $\pi_D = \mathbf{P}\left(F|D\right)$. Esta información se resume en la siguiente tabla de contingencia:

Enfermedad / *F. Riesgo*	**D**	D^c	**Total**
F	n_{11}	n_{12}	n_1
F^c	n_{21}	n_{22}	n_2
Total	n_D	n_C	n

La estimación de la medida es:

(2.6.9.) $$\hat{OR} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}} = \frac{\hat{\pi}_{11} \cdot \hat{\pi}_{22}}{\hat{\pi}_{12} \cdot \hat{\pi}_{21}}$$

Donde $\hat{\pi}_{ij}$; $i, j \in (D,C)$, representa la estimación de la probabilidad π_{ij}; $i, j \in (D,C)$ correspondiente a cada celda . Así, para aquellos enfermos que están expuestos al factor se tiene $\hat{\pi}_{11} = n_{11}/n_D$.

Si existe algún $n_{ij} = 0$, entonces se procede efectuar la corrección dada por:

$$\hat{OR} = \frac{(n_{11} + 0.5) \cdot (n_{22} + 0.5)}{(n_{12} + 0.5) \cdot (n_{21} + 0.5)}$$

En orden a obtener la propiedades del estimador $\hat{OR}$ y un intervalo de confianza para la misma, y dado que no es lineal en $n_{i,j}$, $i, j = (D,C)$, la transformaremos logarítmicamente. Damos una breve reseña sobre la estimación de la varianza de $\log(\hat{OR})$, pues es parte importante para la estimación del IC.

Previamente exponemos algunas cuestiones utilizadas en el desarrollo posterior. Si consideramos la función $log(x)$ y la desarrollamos en el punto $x_0 = 1$, obtenemos: $log(x) = x - 1 + O((x-1)^2)$. Del mismo modo , utilizando *un método delta* se tiene:

$$\log\left(\hat{OR}\right) = \log\left(OR\right) + \frac{\hat{\pi}_{11}}{\pi_{11}} + \frac{\hat{\pi}_{22}}{\pi_{22}} - \frac{\hat{\pi}_{12}}{\pi_{12}} - \frac{\hat{\pi}_{21}}{\pi_{21}} + O_p$$

(véase Apéndice I, para cuestiones sobre infinitésimos).

En nuestro caso:

$$\log\left(\frac{\hat{OR}}{OR}\right)=\log\left(\frac{\hat{\pi}_{11}/\pi_{11}\cdot\hat{\pi}_{22}/\pi_{22}}{\hat{\pi}_{12}/\pi_{12}\cdot\hat{\pi}_{21}/\pi_{21}}\right)=\log\left(\frac{\hat{\pi}_{11}}{\pi_{11}}\right)-\log\left(\frac{\hat{\pi}_{21}}{\pi_{21}}\right)+\log\left(\frac{\hat{\pi}_{22}}{\pi_{22}}\right)-\log\left(\frac{\hat{\pi}_{12}}{\pi_{12}}\right)$$

Desarrollando en serie obtenemos: $\log\left(\frac{\hat{\pi}_{11}}{\pi_{11}}\right)-\log\left(\frac{\hat{\pi}_{21}}{\pi_{21}}\right)\approx\frac{\hat{\pi}_{11}}{\pi_{11}}-\frac{\hat{\pi}_{21}}{\pi_{21}}+O_P$, o lo

que es lo mismo: $\log\left(\frac{\hat{\pi}_{11}}{\pi_{11}}\right)-\log\left(\frac{\hat{\pi}_{21}}{\pi_{21}}\right)=\left\{\frac{1}{\pi_{11}}+\frac{1}{\pi_{21}}\right\}\cdot\hat{\pi}_{11}-\frac{1}{\pi_{21}}$

Luego: $\operatorname{var}\left(\log\left(\frac{\hat{\pi}_{11}}{\pi_{11}}\right)-\log\left(\frac{\hat{\pi}_{21}}{\pi_{21}}\right)\right)=\left\{\frac{1}{\pi_{11}}+\frac{1}{\pi_{21}}\right\}^2\cdot\operatorname{var}(\hat{\pi}_{11})=\frac{1}{\pi_{11}\cdot\pi_{21}\cdot n_D}$

Análogamente $\operatorname{var}\left(\log\left(\frac{\hat{\pi}_{22}}{\pi_{22}}\right)-\log\left(\frac{\hat{\pi}_{12}}{\pi_{12}}\right)\right)=\frac{1}{\pi_{22}\cdot\pi_{12}\cdot n_c}$

Consideradas conjuntamente:

$$\operatorname{var}\left(\log\left(\hat{OR}\right)\right)=\frac{n_D}{n_{11}\cdot n_{21}}+\frac{n_C}{n_{22}\cdot n_{12}}=\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}$$

Como $\log\left(\hat{OR}\right)$ se distribuye normalmente, el intervalo de confianza al nivel 1-α, incorporando la corrección por continuidad, viene dado por la expresión:

(2.6.10.) $$\left[\log(\hat{OR})\pm z_{\alpha/2}\cdot\sqrt{\frac{1}{n_{11}+0.5}+\frac{1}{n_{12}+0.5}+\frac{1}{n_{21}+0.5}+\frac{1}{n_{22}+0.5}}\right]$$

El término z_α es el cuantil 1-α de la distribución normal tipificada N(0,1).

2.6.4. Odd-ratio ajustada: método de *Cochran-Mantel-Haenszel.*

El grado de asociación entre el factor de estudio y la enfermedad pudiera estar afectado por un factor de confusión. En tal caso, conviene obtener una medida de asociación ajustada por el posible factor de confusión. Expondremos a continuación el *método de Cochran-Mantel-Haenszel* (*CMH*).

Consideremos una enfermedad D y un posible factor de riesgo F para la misma. Sea C un hipotético factor de confusión con r niveles. Los datos obtenidos pueden estratificarse en r tablas de contingencia, siendo la k-ésima de la forma:

Enfem. / F. Riesgo	D	D^c	Totales
F	n_{11k}	n_{12k}	$n_{1\square k}$
F^c	n_{21k}	n_{22k}	$n_{2\square k}$
Totales	$n_{\square 1k}$	$n_{\square 2k}$	$n_{\square\square k}$

donde $n_{i\square k}=\sum_j n_{ijk}$, $n_{\square jk}=\sum_i n_{ijk}$ y $n_{\square\square k}=\sum_{i,j} n_{ijk}$, para cualquier valor $k=1,\cdots,r$. La expresión de la odd-ratio ajustada por el método *CMH* tiene la forma:

$$\hat{OR}_{CMH}=\frac{\sum_{k=1}^{r}(n_{11k}\cdot n_{22k}/n_{\square\square k})}{\sum_{k=1}^{r}(n_{12k}\cdot n_{21k}/n_{\square\square k})} \qquad (2.6.11.)$$

La tabla anterior presenta odd-ratios diferentes, definidos por niveles de la variable de confusión; esto es, odd-ratios condicionados. Si la variable de confusión no fuera tal, los odd-ratios condicionados valdrían aproximadamente lo mismo (salvo pequeñas diferencias debidas al azar). Por otro lado si no existiera relación entre los expuestos al factor y la enfermedad los odd-ratios serían iguales a la unidad.

Antes de arriesgarnos a estimar una odd-ratio única empleando la (2.6.11.), debe ser comprobada la homogeneidad de los odd-ratios de las r tablas consideradas, pues efectivamente solo se puede estimar una conjunta si las de las r tablas presentan un cierto grado de homogeneidad. Exponemos brevemente la propuesta de *Breslow and*

Day (1980), conocida como *prueba de homogeneidad de las odd-ratios*; consiste en la comparación entre las frecuencias observadas y los esperados bajo la hipótesis de homogeneidad. El estadístico a evaluar es:

$$\chi^2_{BD} = \sum_k (n_{11k} - \hat{e}_k)^2 / \hat{v}_k$$

donde $\hat{e}_k$, es obtenido como una solución de la ecuación cuadrática:

$$\hat{e}_k = \left(n_{1.k} \cdot \hat{OR}_{CMH}\right) + n_{2.k} \pm \left\{\left[\hat{OR}_{CMH} \cdot \left(n_{1.k} + n_{.1k}\right) + n_{2.k}\right]^2 - \left[4\left(\hat{OR}_{CMH} - 1\right)\hat{OR}_{CMH}\left(n_{1.k} \cdot n_{.1k}\right)\right]\right\}^{1/2} \Big/ 2\left(\hat{OR}_{CMH} - 1\right)$$

la expresión $\hat{v}_k$, estima la varianza, bajo la asunción de homogeneidad:

$$\hat{v}_k = \left(\frac{1}{\hat{e}_k} + \frac{1}{n_{1.k} - \hat{e}_k} + \frac{1}{n_{.1k} - \hat{e}_k} + \frac{1}{n_{2.k} - n_{.1k} + \hat{e}_k}\right)$$

El estadístico χ^2_{BD}, sigue un distribución ji-cuadrado con *r*-1 grados de libertad

Resulta interesante el contraste de hipótesis sobre la independencia del factor y la enfermedad.

H_0: *No existe relación entre factor y enfermedad.*

H_1: *Factor y enfermedad están asociados.*

Bajo la hipótesis nula, se tiene:

$$E[n_{11k}] = \frac{n_{1\square k} \cdot n_{\square 1k}}{n_{\square\square k}}$$

$$\text{var}(n_{11k}) = \frac{n_{1\square k} \cdot n_{2\square k} \cdot n_{\square 1k} \cdot n_{\square 2k}}{n^2_{\square\square k} \cdot (n_{\square\square k} - 1)}$$

Apoyándonos en estas dos expresiones, calculamos el estadístico de contraste, llamado de *Cochran-Mantel-Haenszel*:

$$CMH = \frac{\left[\sum_k (n_{11k} - E[n_{11k}])\right]^2}{\sum_k \text{var}(n_{11k})}$$

Se demuestra que este estadístico, para muestras grandes, sigue una distribución ji-cuadrado con un grado de libertad; por lo tanto solo aceptamos la hipótesis alternativa *si* $CMH > \chi^2$ (1) ($\alpha = 0.05$) , y con un p-valor<0.05. Un intervalo de confianza, al nivel $1-\alpha$, para este estimador de odd-ratio $\hat{OR}$ viene dado por la siguiente expresión:

$$\hat{OR}_{CMH} = 1 \pm z_\alpha \left\{ \sum_k \text{var}(n_{11k}) \right\}^{1/2}$$

En determinadas condiciones, abundancia de datos repartidos homogéneamente a través de las r tablas, la estimación $\hat{OR}_{CMH}$ es equivalente a la odd-ratio ajustada por el modelo *logit* (véase epígrafe 2.9.) Estas son las condiciones de desarrollo de esta memoria; por lo tanto, en el capítulo 3 (Estudio de hipertensión arterial en Lanzarote) solamente se estimarán las odd-ratios ajustadas por el modelo *logit*.

2.7. Discriminación univariante: curvas ROC.

Como parte del estudio de hipertensión arterial, se consideran marcadores univariantes, con poder de discriminación que nos permitan clasificar en grupos (normotenso e hipertenso, en nuestro caso) ; cuestión que pasa por analizar la bondad de los protocolos diagnósticos basados en factores de riesgo. Esto se lleva a cabo mediante la estimación de curvas *ROC* (*Receiver Operating Characteristics*). Esta metodología construye una curva basándose en pares, para cada valor de la variable utilizada para el diagnóstico (*falsos positivos, sensibilidad)*. Podemos calcular puntos *cut-off* donde, para un determinado valor de la variable, se obtienen valores de sensibilidad y especificidad idóneos para los objetivos perseguidos. Se implementan distintos criterios

de determinación de puntos *cut-off*, que responden a las distintas interrogantes planteadas, según las diferentes necesidades.

Se consideran varias metodologías de estimación de probabilidades, aplicadas a la construcción de curvas *ROC*: una estimación *empírica cruda*, otra basada en hipótesis de *normalidad* y por última una *estimación no paramétrica.* Esta última metodología es objeto de especial exposición.

Consideremos una población en la que existe una determinada enfermedad. Se conoce el hecho de que esa enfermedad tiene un factor de riesgo X, que utilizamos como marcador numérico. Los enfermos seguirán una distribución de probabilidad $F_D(x)$ y los no afectados por esa enfermedad (sanos), tendrán otra distribución de probabilidad $F_{D^c}(x)$.

Supongamos, sin pérdida de generalidad, que la enfermedad provoca una disminución en el marcador X. Lo que se busca es un valor del marcador (*cut-off*) ,C, tal que para $X \leq C$, diagnostique que el individuo padece la enfermedad. De modo inverso, pero metodológicamente análogo, procederemos si la patología es de aumento. Brevemente exponemos varios conceptos que resultan imprescindibles para los siguientes desarrollos:

Sensibilidad: *es un coeficiente que, para un determinado protocolo diagnóstico y un valor del marcador, mide la capacidad del protocolo, para detectar la enfermedad en individuos, que realmente están enfermos.*

Expresado en términos de probabilidad: $\mathbf{P}(+|D)$ Donde D es el suceso padecer la enfermedad y "+" haber dado positivo en el protocolo diagnóstico. Considerando que la enfermedad disminuya el marcador, en los mismos términos expresados más arriba, esta definición se expresa de la siguiente manera:

$$s = \mathbf{P}(X \leq C|D)$$

Falso positivo: *es un coeficiente que, para un determinado protocolo diagnóstico y un valor del marcador, mide la capacidad del protocolo para detectar la enfermedad en individuos, que realmente no la padecen.*

Expresado en términos de probabilidad: $\mathrm{P}\left(+\middle|D^c\right)$. Donde D es el suceso padecer la enfermedad y "+" haber dado positivo en el protocolo diagnóstico. Considerando que la enfermedad disminuya el marcador, en los términos anteriores, esta definición se expresa de la siguiente manera:

$$\varphi = \mathbf{P}\left(X \leq C \middle| D^c\right)$$

Especificidad *como la capacidad de un protocolo diagnóstico, considerado un valor determinado del marcador, para descartar la enfermedad en individuos realmente sanos.* Su valor $\varepsilon = 1 - \varphi$.

La curva ROC viene defina por la función $s = s(\varphi)$. La bondad del protocolo diagnóstico viene estimado en términos de área bajo la curva. Valores próximos a la unidad indican que el protocolo es eficiente; si el valor del área está próximo a 0.5 el protocolo produce un pobre diagnóstico.

Expresadas en términos de función de distribución, para el valor del cut-off C, se obtiene: $s = F_D(C)$ y $\varphi = F_{D^c}(C)$. Asumiendo que las funciones de distribución F_D y F_{D^c}, son continuas y estrictamente crecientes, existen sus inversas, de tal modo que: $C = F_D^{-1}(s)$ y $C = F_{D^c}^{-1}(s)$, igualando estas dos expresiones de C, se obtiene: $s = F_D\left(F_{D^c}^{-1}(\varphi)\right)$, expresión gráfica de la curva ROC en función de ambas funciones de distribución.

Señalemos que si la patología se traduce en un aumento del marcador X, el tratamiento es análogo al anterior, sin más que declarar enfermos a los individuos tal que su marcador sea $X \geq C$. Evidentemente los coeficientes Falso Positivo y Sensibilidad, en este caso son: $1 - F_{D^c}(x)$ y $1 - F_D(x)$.

Cuando los datos no se ajustan a la distribución normal o ésta se desconoce, las funciones de distribución podrían estimarse por métodos no paramétricos. Consideremos por tanto una muestra aleatoria $X_1,\ldots,X_n$ de una función de distribución de probabilidad $F(x)$. El estimador de núcleo de $F(x)$ se define por:

$$\hat{F}_n(x,h)=\frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x-X_i}{h}\right) \qquad (2.7.1.)$$

siendo $W(x)$ una función de distribución y h el parámetro de alisamiento o bandwidth. Esta función a menudo recibe el nombre de núcleo integrado, dado que su derivada es una función de núcleo en el sentido ordinario (función de densidad, simétrica, continua y de soporte compacto).

Estas ideas sobre estimación no paramétrica fueron introducidas por Rosenblatt (1956) y posteriormente completadas y desarrolladas por Azzalini (1981). En los métodos para la obtención del parámetro de alisamiento o bandwitch óptimo, tienen especial incidencia las aportaciones de Hall (1992), Bowman and Azzalini (1997), Bowman , Hall and Pravant (1998) y Härdle (1999).

Consideraremos núcleos integrados $W(x)$, siendo $W(x)=\int_{-\infty}^{x} k(u)\,du$ y $K(u)$ una función de núcleo tal que $K(x)=W'(x)$ es una función lipschitziana, de soporte compacto, continua y con momento de segundo orden finito. En esta memoria se utilizará el *núcleo de Epanechnikov*, definido por $K(u)=(3/4)(1-u^2)I(|u|\le 1)$, lo que supone que el núcleo integrado tiene la forma $W(z)=\int_{-1}^{z} 3(1-x^2)/4\,dx=(2+3z-z^3)/4$.

Para estimar el bandwidth óptimo usado en la función de distribución $\hat{F}_n(x;h)$, se debe calcular la varianza y el sesgo (bias) del estimador. Suponemos $W(x)$ derivable, con $K(x)=W'(x)$ donde $K(x)$ es un núcleo y $F\in C^2$. Entonces:

(2.7.2.)

$$\text{i)}\ E\left[\hat{F}_n(x;h)\right] - F(x) = \frac{F''(x)\mu_2(K)}{2}h^2 + o(h^2),\quad h \to 0$$

$$\text{ii)}\ \operatorname{var}\left(\hat{F}_n(x;h)\right) = \frac{F(x)(1-F(x))}{n} - \frac{F'(x)\mu_1\left(\left(W^2\right)_\bullet\right)}{n}h + o(h),\ h \to 0$$

siendo $\mu_2(K) = \int x^2 K(x)dx$ y $\mu_1\left(\left(W^2\right)_\bullet\right) = \int y \cdot (\partial/\partial y) W^2(y)dy$

Considerando las expresiones (2.7.1.) que representan sesgo y varianza respectivamente de $\hat{F}_n(x;h)$, observamos que la primera es proporcional a h^2, y la varianza lo es a $(nh)^{-1}$; lo que significa que para minimizar el sesgo, se debe elegir un *h pequeño*, pero esta opción tiende a maximizar la varianza. Para conciliar estas dos opciones definimos el *MISE* (Mean integral square error), que nos mide la bondad de la estimación en términos de varianza y bias. Viene dado por:

$$\text{MISE}(h) = E\left[\int \left\{\hat{F}_n(x,h) - F(x)\right\}^2 dx\right]$$

que es igual a:

$$\text{MISE}(h) = \int \operatorname{var}\left(\hat{F}_n(x;h)\right)dx + \int \left\{E\left[\hat{F}_n(x;h)\right] - F(x)\right\}^2 dx;$$

cuya expresión desarrollada es:

(2.7.3.)

$$\text{MISE}(\text{h}) = \frac{1}{nh} \| F'' \|_2^2 + \frac{h^4}{4}(\mu_2(K))^2 \ \| f''(x) \|_2^2 + o((nh)^{-1}) + o(h^4),$$

$$h \to 0 \ \ y \ \ nh \to \infty$$

donde $\|F''\|_2^2 = \int F''(x)^2 dx$.

En las condiciones de la (2.7.3.) *MISE* converge a cero y el bandwidth óptimo viene definido por: $h_0 = \left\{\frac{\mu_1\left(W_\bullet^2\right)}{\mu_2(K)^2 \cdot \|F''\|_2^2 \cdot n}\right\}^{1/3}$

Sin embargo esta expresión (2.7.3.) es muy poco operativa. Para salvar esta dificultad se exponen a continuación dos métodos. El primero se conoce como *Regla del Pulgar (RP)*, y es una estimación poco afinada de *h*. El segundo es el de la *Máxima Verosimilitud (MV)* (validación cruzada), mucho más costoso en proceso, pero más exacto en cuanto al valor del bandwidth óptimo *h*.

La Regla del Pulgar es una estimación basada en el supuesto de normalidad; h_0 viene dada por:

$h_0 = 0.75 \text{ min.} \left(s, \frac{\hat{R}}{\lambda} \right) . n^{-1/5}$. Donde s es la desviación típica muestral.

$\hat{R}$ =Rango Intercuartilico. *n*=tamaño muestral y λ = máx. (λ_D, λ_C), con $\lambda_c = \frac{\hat{R}_c}{s_c}$ en referencia a los estadísticos de controles; y $\lambda_D = \frac{\hat{R}_D}{s_D}$ en referencia a casos.

La segunda opción (MV), consideramos una serie de observaciones X_i, independientes (adicionales sobre las que ya tenemos). La máxima verosimilitud para esas observaciones viene dada por $\prod_i \hat{F}'_h(X_i)$, el valor de este estadístico para los diferentes valores de *h*, nos indicará que bandwidth es preferible. Cuando no se dispone de otras observaciones se opta por dejar una fuera cada vez; esto es equivalente a considerar el conjunto de observaciones como $\{X_j\}_{i \neq j}$, así la función de densidad global estimada será:

$$\prod_{i=1}^{n} \hat{F}'_{h,i}(X_i) = (n-1)h^{-n} \prod_{i=1}^{n} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h} \right)$$

Tomando el logaritmo y normalizando mediante el factor n^{-1}, se obtiene:

$$ML\ (h) = n^{-1} \sum_{i=1}^{n} \log\left[\hat{F}'_{h,i}(X_i) \right] = n^{-1} \sum_{i=1}^{n} \log\left[\sum_{j \neq i} K\left(\frac{X_i - X_j}{h} \right) \right] - \log\left[(n-1)h \right]$$

El bandwidtch óptimo vendrá dado por el máximo de esta función: $h_0 = \max_h (ML\ (h))$

2.7.1. Diversos criterios para elección de puntos Cut-off.

En el epígrafe 2.7 de esta memoria ya se ha descrito la curva ROC, y en particular los puntos cut-off. Nos proponemos dar diversos criterios para su determinación, que responden a distintos intereses de conocimiento sobre la información suministrada por un protocolo diagnóstico.

- ***Criterio de Máxima Sensibilidad + Especificidad***: el cut-off correspondiente, señala, conjuntamente para enfermos y para sanos, el punto de óptimo comportamiento de la curva. En él se alcanza el máximo de sensibilidad más especificidad. Proporciona un criterio equilibrado entre ambos conceptos.

  $$C_1 = \max_i(s(x_i) + \varepsilon(x_i)).\ \forall x_i \text{ valor del marcador numérico } X.$$

- ***Criterio de Máxima Sensibilidad***: respetando el criterio anterior, se exige una sensibilidad igual o superior al 90%, o valor inferior más próximo.

  $C_2 = \max_i(\gamma + \varepsilon(x_i))$. Con $\gamma \geq 0.90$, o bien $\gamma = \max_i s(x_i)$, $\forall x_i$ valor del marcador numérico X.

  Evidentemente, como en Estadística no obtenemos algo a cambio de nada, el alto grado de sensibilidad se consigue, a cambio de sacrificar especificidad. Suele proporcionar un buen cut-off para detectar la enfermedad, pero mediocre para descartarla. En algún caso podría coincidir con el anterior criterio.

- ***Criterio de intervalo basado en VPN y VPP***: En los dos criterios anteriores buscábamos un único punto cut-off, que respondiese a nuestra demanda de información. Sin embargo, podemos considerar no un único cut-off, sino un intervalo $[C_3, C_4]$, tal que, si el protocolo diagnóstico es bueno, y, llevando la patología asociado un aumento del marcador, declararía enfermos, con una altísima probabilidad a aquellos cuyo marcador es $X > C_4$; y declararía sanos, con una también gran probabilidad a aquellos cuyo marcador $X < C_3$. Si el marcador de un individuo pertenece al interior del intervalo, se le practicarán pruebas complementarias. Si la patología conllevara una disminución del marcador, bastaría cambiar las desigualdades.

Si la patología, sin perdida de generalidad, aumenta el valor del marcador, el extremo inferior del intervalo, C_3, viene definido por *VPN* (Valor Predictivo Negativo), cuya expresión es la siguiente:

$VPN(C) = \mathbf{P}\left(D^c \,|\, X < C\right)$. El extremo superior , C_4, viene definido por *VPP* (Valor Predictivo Positivo), que se expresa como:

$$VPP(C) = \mathbf{P}\left(D \,|\, X > C\right).$$

Desarrollando ambas expresiones por el teorema de Bayes:

$$VPN(C) = \frac{\mathbf{P}\left(X < C \,|\, D^c\right)\mathbf{P}\left(< D^c\right)}{\mathbf{P}\left(X < C \,|\, D^c\right)\mathbf{P}\left(D^c\right) + \mathbf{P}\left(X < C \,|\, D\right)\mathbf{P}\left(D\right)} = \frac{\varepsilon \cdot (1-\pi)}{\varepsilon \cdot (1-\pi) + \beta \cdot \pi}$$

Donde ε=Especificidad; π=Prevalencia de la enfermedad; β=Falsos negativos.

$$VPP(C) = \frac{\mathbf{P}\left(X > C \,|\, D\right)\mathbf{P}(D)}{\mathbf{P}\left(X > C \,|\, D\right)\mathbf{P}(D) + \mathbf{P}\left(X > C \,|\, D^c\right)\mathbf{P}(D^c)} = \frac{s \cdot \pi}{s.\pi + \gamma \cdot (1-\pi)}$$

Donde s = Sensibilidad; γ=Falsos positivos.

Para este criterio podemos elegir $C_3 = \max_i VPN(X_i)$ y $C_4 = \max_i VPP(X_i)$. O bien exigiendo, para VPP y VPN, sendos valores cut-off, tal que se mantengan por encima de valores mínimos; orientativamente $C_3 = 0.80$ y $C_4 = 0.95$.

2.7.2. Simulación.

Se ha procesado una muestra de tamaño $n = 489$, donde únicamente figura una variable, de rango [1, 153], que no siguen una distribución normal. Para cada caso, se conoce su estado (sano o enfermo). Se han considerado dos supuestos. El primero (color blanco en Tabla 2.7.4.) considera que el protocolo diagnóstico eficiente (Área bajo la curva ROC: orden de 90%). En el segundo caso (color gris en Tabla 2.7.4.) se considera un protocolo diagnóstico menos eficiente (Área bajo curva ROC: orden 70%). Ambas alternativas proceden de la misma base de datos, ligeramente alterada. En la figura 2.7.5. se muestran las curvas correspondientes a los tipos de diagnósticos, así como los dos tipos de estimaciones (cruda y no paramétrica, esta última con dos estimaciones diferentes del bandwidth).

Tabla 2.7.4. Simulación curvas ROC.

RESULTADOS SIMULACIÓN.										
DIAGNÓSTICO EFICIENTE. ($n = 489$, PREVALENCIA π=39.7%)										
MÁX(ESP+SENSI)			MÁX SEN			INTV. VPN.		INTV. VPP.		ÁREA BAJO CURVA ROC.
C_1	s	ε	C_2	s	ε	C_3 (Prob.)		C_4 (Prob.)		
ESTIMACIÓN CRUDA.										
21	0.855	0.932	18	0.90	0.87	1	(1)	48	(0.95)	0.927
ESTIMACIÓN NO PARAMÉTRICA. (R.P. $h_D = 5.74, h_C = 2.15$)										
19	0.88	0.88	17	0.91	0.82	1	(0.91)	42	(0.93)	0.877
ESTIMACIÓN NO PARAMÉTRICA. (M.V. $h_D = 1.57, h_C = 1.034$)										
19	0.88	0.89	18	0.90	0.86	1	(0.99)	49	(0.93)	0.896
DIAGNÓSTICO MENOS EFICIENTE. ($n = 489$, PREVALENCIA π=41.1%)										
MÁX(ESP+SENSI)			MÁX SEN			INTV. VPN.		INTV. VPP.		ÁREA BAJO CURVA ROC.
C_1	s	ε	C_2	s	ε	C_3 (Prob.)		C_4 (Prob.)		
ESTIMACIÓN CRUDA.										
17	0.68	0.69	5	0.91	0.26	1	(1)	90	(0.80)	0.709
ESTIMACIÓN NO PARAMÉTRICA. (R.P. $h_D = 5.82, h_C = 4.006$)										
18	0.65	0.69	1	0.97	0.06	1	(0.99)	58	(0.68)	0.639
ESTIMACIÓN NO PARAMÉTRICA. (M.V. $h_D = 1.51, h_C = 1.059$)										
17	0.66	0.69	5	0.91	0.23	1	(0.83)	60	(0.69)	0.678

Sobre el supuesto de diagnóstico eficiente, Tabla 2.7.4 (parte en color blanco), cabe destacar que en este protocolo es muy bueno, pues se obtienen áreas bajo la curva entre 0.92 7 y 0.89. Obsérvese que la estimación no paramétrica, cuyo bandwidth está basado en Máxima Verosimilitud (M-V), proporciona resultados más próximos a la

estimación cruda, que la basada en la Regla del pulgar (R-P); en términos de área bajo la curva se obtiene 0.896, frente al 0.877 obtenido por el segundo método. También en cuanto a los punto cut-off los valores y sus probabilidades están más próximos con M-V, que con R-P.

Cabe observar que los parámetros de alisamiento, se obtienen más refinados con M-V. que con R-P, pues sus valores son respectivamente 1.567 para los enfermos , 1.034 para los sanos y 5.737, 2.149 para R-P. Por construcción ambas estimaciones del bandwidth óptimo, cumplen la expresión (2.7.3.); si bien, con respecto a la primera, las convergencias $h \to 0$ y $nh \to \infty$, son más rápidas con M-V, que con R-P. Sobre la estimación realizada en el supuesto menos eficiente, los resultados son más pobres (el área bajo la curva está entre 0.705 y 0.639 para R-P), siendo de aplicación todo lo dicho para el caso de eficiencia.

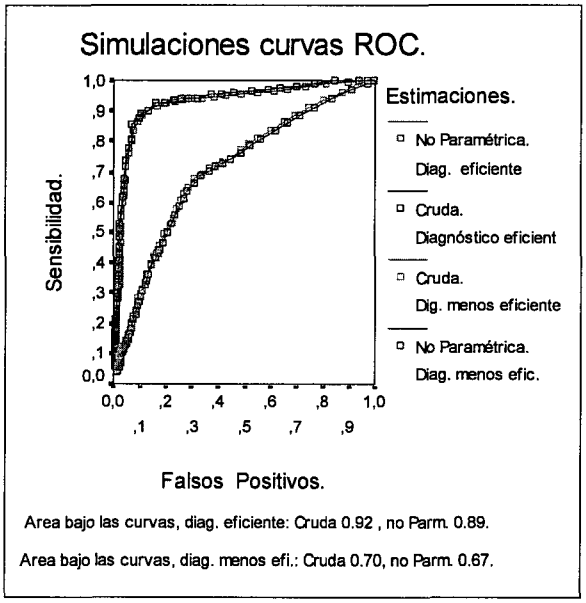


Fig 2.7.5.

Por último la figura 2.7.5, muestra una comparativa de las curvas ROC para los dos supuestos, siendo de destacar la buena estimación que proporcionan los métodos no paramétricos, en especial M-V, que es el representado. Es evidente la diferencia que existe, en términos de área bajo la curva y forma de la misma, entre los dos supuestos diagnósticos.

2.8. Marcadores multidimensionales: análisis discriminante.

A diferencia del epígrafe 2.7 en que usábamos un único marcador, mostramos en esta sección la posibilidad de construir marcadores para una clasificación dada, basados en un conjunto de variables, de tal forma que los nuevos marcadores posean una mayor potencia discriminadora que cada uno a nivel individual (véase epígrafe 2.7.2.). Para ilustrar esta idea, obsérvese la figura 2.8.1.

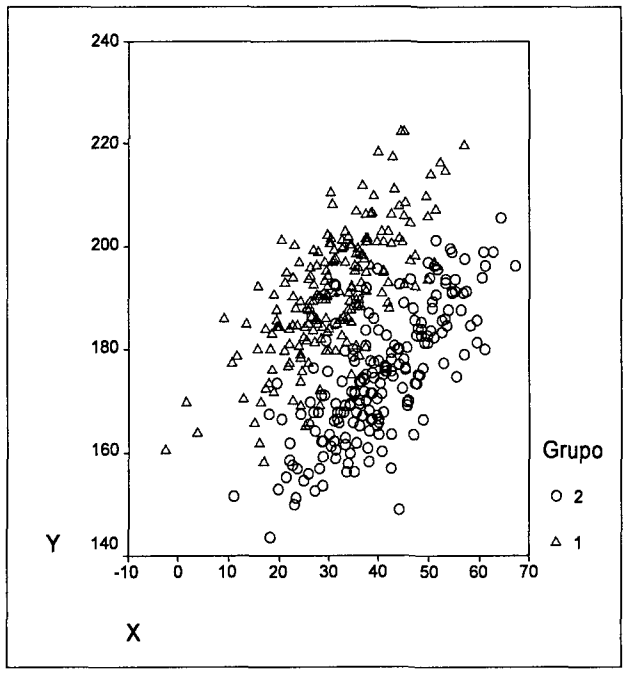


Fig. 2.8.1.
Representación simultánea de dos variables X e Y según el grupo de pertenencia correspondiente a una clasificación dada.

En esta gráfica puede observarse que las variables X e Y discriminan de forma muy evidente entre los dos grupos considerados. Sin embargo, si se consideran las proyecciones de los puntos sobre los ejes coordenados (las proyecciones sobre el eje de abscisas son los valores de la variable X, y sobre el eje de ordenadas, los valores de la variable Y), la discriminación es casi imperceptible. Esto puede comprobarse a través de las curvas ROC correspondientes a cada una de las variables consideradas figura 2.8.2.

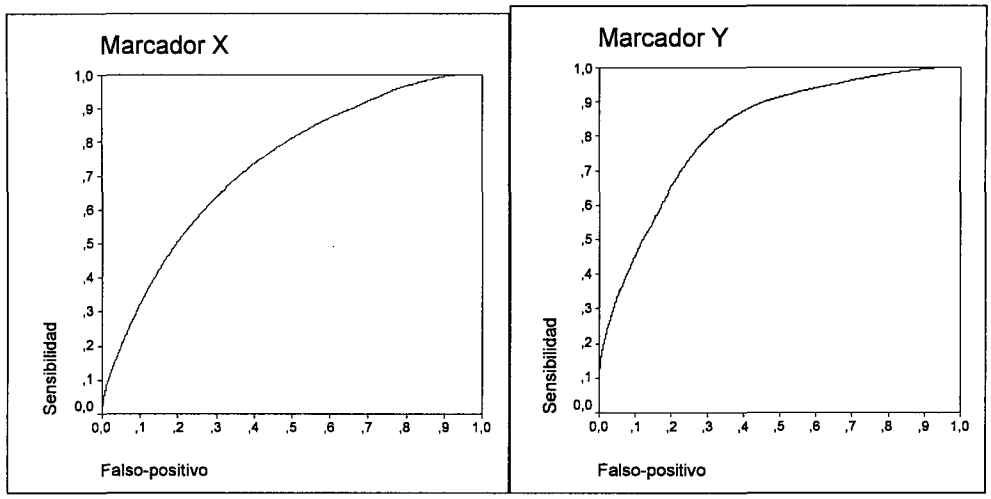


Fig. 2.8.2. Curvas ROC. (a) Para el marcador *X*. Área bajo la curva 0.73. (b) Marcador *Y*. Área bajo la curva 0.81

2.8.1. Funciones discriminantes canónicas.

El objetivo de los marcadores multivariables es obtener una función (score) construido a partir de varias variables, que acumule la información discriminante que dan a nivel individual. A través de las técnicas del análisis discriminante se ha determinado un marcador *Z*, el cual es una función lineal de *X* e *Y*; esto es:

$$Z = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot Y$$

Si en un individuo se pueden determinar los valores de las variables *X* e *Y*, también podrá determinarse el valor del score *Z*. Esta variable acumula la capacidad discriminadora contenida en las variables originales *X* e *Y*. Los valores de los coeficientes de la función lineal β_0, β_1 y β_2 se obtienen a través de los procedimientos del análisis discriminante. La figura 2.8.3. muestra la curva ROC correspondiente al score Z.

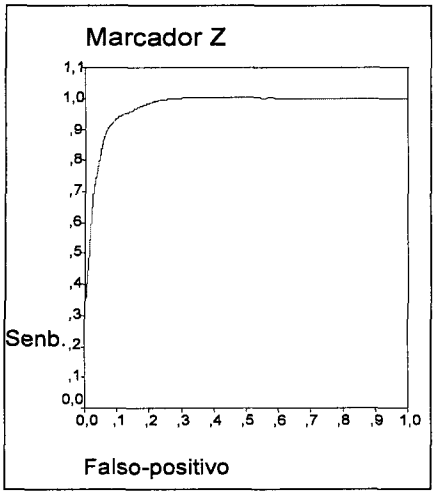


Fig. 2.8.3.

Curva ROC para el marcador Z. Área bajo la curva 0.97

El estudio expuesto es una simulación en la que se ha querido mostrar cómo variables que a nivel individual tienen una escasa capacidad discriminadora, conjuntamente dan una capacidad muy elevada.

Consideremos una clasificación $C_1,\cdots,C_m$ de un conjunto $\Omega(\bigcup_{i=1}^{m} C_i = \Omega$ y $C_i \bigcap C_j = \varnothing, i \neq j)$. Sea $X_1,\cdots,X_p$ un vector de variables evaluadas sobre los elementos de □ con el objetivo de discriminar entre las clases dadas. Representamos por x_{ijk} el valor de la *k*-ésima variable evaluada sobre el *j*-ésimo objeto perteneciente a la *i*-ésima clase. De esta forma, el conjunto de datos para la realización del análisis discriminante tiene la forma:

$$\left\{ x_{ijk} ; i = 1,\cdots,m; j = 1,\cdots,n_i; k = 1,\cdots,p \right\}$$

La base de datos puede representarse como:

Clase	X_1	...	X_j	....	X_p
1	x_{111}		x_{11j}		x_{11p}
	.....				
	$x_{1n_1 1}$		$x_{1n_1 j}$		$x_{1n_1 p}$
...	.....				
i	x_{i11}		x_{i1j}		x_{1ip}
	.....				
	$x_{in_i 1}$		$x_{in_i j}$		$x_{in_i p}$

(Continua.)

...	.....		
m	x_{m11}	x_{m1j}	x_{m1p}
	.....		
	x_{mn_m1}	x_{mn_mj}	x_{mn_mp}

Tabla 2.8.4.

Una variable canónica discriminante es una combinación lineal de las variables observadas de la forma:

$$Z = \beta_1 X_1 + \cdots + \beta_p X_p$$

Los coeficientes $\beta_1, \cdots, \beta_p$ pueden elegirse de tal forma que la variable Z maximice la discriminación entre las clases $C_1, \cdots, C_m$. Esto es, que su distribución tenga variación máxima sobre las clases consideradas. Un criterio para hacerlo es maximizar el F-test del análisis de la varianza. De esta forma es posible construir s variables canónicas $Z_1, \cdots, Z_s$ de la forma:

(2.8.5.) $$Z_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \; ; i = 1, \cdots, s$$

donde:

i. $s = \min\{m-1, p\}$

ii. $Z_1, \cdots, Z_s$ son incorreladas dentro de los grupos

Así, para discriminar entre sólo dos clases, sólo es posible construir una única variable canónica.

2.8.2. Funciones de clasificación lineal de Fisher.

Consideraremos nuevamente el conjunto de clases $C_1,\cdots,C_m$ y el vector de variables discriminantes $\mathbf{X}=\left(X_1,\cdots,X_p\right)$. El objetivo ahora es, dado un objeto arbitrario $\omega\in\Omega$, buscar la clase más probable para este objeto basándose en la evaluación $\mathbf{x}=\mathbf{X}(\omega)$ del vector de variables.

La construcción de las funciones de clasificación lineal de Fisher se basa en la hipótesis de que el vector $\mathbf{X}$ sobre la clase C_i tiene distribución normal multivariable $N_p\left(\mu_i,\boldsymbol{\Sigma}\right)$ (nótese la hipótesis de homoscedasticidad). Un criterio de asignación consistiría entonces es clasificar al objeto w en aquella clase C_{i_0} tal que:

$$f_{i_0}(\mathbf{x})=\max_{1\le i\le m} f_i(\mathbf{x})$$

De esta condición, para cada clase C_i podemos obtener una función lineal de la forma:

(2.8.6.) $$\varphi_i(\boldsymbol{x})=\mu_i\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^t-\frac{1}{2}\mu_i\boldsymbol{\Sigma}^{-1}\mu_i^t\ ;\qquad i=1,\cdots,m$$

La condición anterior es entonces equivalente a lo siguiente: el objeto w pertenece a aquella clase C_{i_0} tal que:

$$\varphi_{i_0}(\mathbf{x})\ge\varphi_i(\mathbf{x});\ \forall i=1,\cdots,m$$

2.8.3. Valores predictivos.

La regla de clasificación basada en las funciones discriminantes lineales de Fisher es tal que, cuando $\omega\in C_{i_0}$, $f_{i_0}(\mathbf{x})\ge f_i(\mathbf{x})$ para todo $i=1,\cdots,m$. Ahora bien, si ocurre esto último, ¿cuánto de probable es que $\omega\in C_{i_0}$?. En definitiva, cuál es el valor de la probabilidad *a posteriori*:

$$\mathbf{P}\left(\omega \in C_{i_0} \middle| \mathbf{x}\right)$$

Nuevamente se requiere disponer de una distribución de probabilidad a priori sobre las clases para evaluar esta probabilidad. Supongamos pues que se pueda asumir que $\mathbf{P}\left(\omega \in C_i\right) = \pi_i$; $i = 1, \cdots, m$. En este caso, de acuerdo con la fórmula de Bayes, la probabilidad a posteriori se calcula por:

$$\mathbf{P}\left(\omega \in C_{i_0} \middle| \mathbf{x}\right) = \frac{f_{i_0}\left(\mathbf{x}\right) \cdot \pi_{i_0}}{\sum_{i=1}^{m} f_i\left(\mathbf{x}\right) \cdot \pi_i}$$

Esta regla, como veremos, puede ser evaluada mediante un procedimiento de validación cruzada. Más concretamente, si disponemos de un conjunto de objetos ya clasificados, estos pueden ser reclasificados aplicando la regla anterior. De esta forma podría estimarse la sensibilidad de la prueba.

$$\mathbf{P}\left(\omega \in C_{i_0} \middle| \mathbf{x}\right) = \frac{f_{i_0}\left(\mathbf{x}\right) \cdot \pi_{i_0}}{\sum_{j=1}^{m} f_j\left(\mathbf{x}\right) \cdot \pi_j} = \frac{f_{i_0}\left(\mathbf{x}\right) \cdot \pi}{\sum_{j=1}^{m} f_j\left(\mathbf{x}\right) \cdot \pi} \geq \frac{f_i\left(\mathbf{x}\right) \cdot \pi}{\sum_{j=1}^{m} f_j\left(\mathbf{x}\right) \cdot \pi} = \mathbf{P}\left(\omega \in C_i \middle| \mathbf{x}\right)$$

En realidad, $f_i\left(\mathbf{x}\right) > f_j\left(\mathbf{x}\right)$ significa que los datos hacen más verosímil la pertenencia de $\omega_{\cdot}$ a la clase C_i que a la C_j. Sin embargo, este hecho no tiene en cuenta el criterio subjetivo del observador que se expresa a través de las probabilidades a priori. La regla de clasificación bayesiana da mayor probabilidad de pertenencia a la clase C_i que a la C_j cuando:

$$f_i\left(\mathbf{x}\right) \cdot \pi_i > f_j\left(\mathbf{x}\right) \cdot \pi_j$$

De la desigualdad anterior se deduce que $\mathbf{P}\left(\omega \in C_i \middle| \mathbf{x}\right) > \mathbf{P}\left(\omega \in C_j \middle| \mathbf{x}\right)$.

2.8.4. Selección de variables para la discriminación.

Hasta ahora a partir de un conjunto de variables hemos determinado funciones discriminantes de éstas. El problema que tratamos en esta sección es determinar precisamente qué variables son realmente aptas para la construcción de funciones discriminantes. Más concretamente, dado un conjunto de variables, se trata de determinar un subconjunto óptimo de variables discriminantes.
Supongamos por tanto que tenemos un conjunto de datos de la forma:

$$\left\{X_{ijk}\,;\, i=1,\cdots,m;\, j=1,\cdots,n_i;\, k=1,\cdots,p\right\}$$

Obsérvese que este conjunto es el resultado de determinar p variables sobre un conjunto de objetos clasificados en m clases. Proponemos ahora un procedimiento para seleccionar un subconjunto de r variables ($r \le m$) a partir del cual obtener las funciones discriminantes. El método consiste en ir sucesivamente incorporando variables utilizando los F-test del análisis de la varianza. Para la k-ésima variable X_{ijk}, consideramos el modelo de análisis de la varianza $X_{ijk} \cong N\left(\mu_{ik};\sigma\right)$; $j=1,\cdots,n_i$. Nótese que para cada variable (k) y en cada clase (i), el centro de gravedad de la distribución es μ_{ik}. El algoritmo de selección que proponemos es:

Paso 1. Para la k-ésima variable, consideramos el F-test para el contraste $H_0 : \mu_{ik} = \mu_k \,;\forall i$ definido por:

$$F_k = \frac{\frac{1}{m-1}\sum_{i=1}^{m} n_i \left(\bar{X}_{i\square k} - \bar{X}_{\square\square k}\right)^2}{\frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(X_{ijk} - \bar{X}_{i\square k}\right)^2}$$

siendo $N = \sum_{i=1}^{m} n_i$, $\bar{X}_{i\square k} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ijk}$ y $\bar{X}_{\square\square k} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ijk}$

Paso 2. Se selecciona como variable discriminante aquella variable para la cual sea máximo F_k siempre que éste sea significativo ($p \le .05$).

Paso 3. Una vez que se han introducido l variables, para cada variable (s) no seleccionada se considera el modelo de análisis de la covarianza:

$$X_{ijs} \cong N\left(\mu_{is} + \sum_{k=1}^{l} \beta_{sk} \cdot X_{ijk}, \sigma\right)$$

y se obtienen los F-test F_s para los contrastes $H_0 : \mu_{is} = \mu_s \; ; \; \forall i$

Paso 4. Se introduce en el conjunto de variables discriminantes la que tenga el mayor valor F_s siempre que éste sea significativo ($p \leq .05$).

Paso 5. El algoritmo se detiene cuando en algún paso no haya ningún F-test significativo o todas las variables hayan sido introducidas en el análisis.

2.9. Regresión logística.

El uso de la metodología de regresión logística en esta memoria tiene como finalidad obtener odd-ratios entre variables binarias, ajustadas por un conjunto arbitrario de variables. En esta sección se realiza una revisión de aquellos aspectos de la metodología que se utilizarán en el estudio de hipertensión arterial que nos ocupa. Se introduce en primer lugar la forma del modelo para variable de respuesta binaria, analizándose el significado de sus parámetros. Se analiza el método de estimación de máxima verosimilitud, estableciéndose condiciones bajo las cuales el estimador de máxima verosimilitud existe y es único. Se exponen estrategias de construcción de modelos y se desarrolla un algoritmo de selección prospectiva de variables. Se revisan asimismo los tradicionales contrastes de bondad ajuste y se introduce el test de Hosmer y Lemeshow, específico para los modelos logit. Un aspecto importante en la construcción de modelos de regresión es la detección de las llamadas observaciones influyentes. Este método se muestra en esta sección para la regresión logística. Para aquellos conjuntos de datos en los que el número de variables predictoras es *grande* en

relación con el número de casos, las estimaciones de los parámetros pueden tener sesgos largos. Para corregir este problema, se introduce una metodología debida a Steyerber (2000) para las correcciones del sesgo basada en el bootstrap.

Esta revisión se realiza en el contexto de diseños en los que, para un vector de variables predictoras $\mathbf{x}$, se observa una variable aleatoria binaria Y. Sin embargo, el diseño del estudio que nos ocupa en esta memoria es de caso-control, lo que supone que la referida variable binaria Y es realmente la variable por la que se realiza la estratificación. Sin embargo, como se pone de manifiesto en 2.9.7 los resultados para las estimaciones de las odd-ratios son igualmente válidos para los diseños de caso-control.

2.9.1. Generalidades sobre los modelos de regresión logística.

Los modelos de regresión logística fueron introducidos por Cox (1970) con el objetivo de describir la dependencia de una variable binaria de un conjunto de variables continuas. No obstante, pueden también utilizarse para describir la asociación entre la referida variable binaria y un conjunto de variables que pueden ser continuas y categóricas, sustituyendo cada variable categórica por un subconjunto de variables dummies asociadas a ella como se mostrará posteriormente. Asimismo, la variable binaria de respuesta puede extenderse también a una variable ordinal, e incluso, a una variable categórica. No obstante, para los fines de esta memoria consideramos únicamente variables de respuesta binaria.

Revisamos pues en esta sección las generalidades de los modelos de regresión logística con variable de respuesta binaria. Tales modelos revisten especificidades en los estudios de caso control que posteriormente contemplaremos. Consideremos por tanto una variable binaria Y que en general representa la presencia o ausencia de una característica de interés y cuyos valores codificamos como (1=*Presencia*) y (0=*Ausencia*). El objetivo inicial es explicar la variable Y por un conjunto de predictores $X_1,\cdots,X_p$, inicialmente numéricos, aunque como posteriormente veremos, podrán considerarse también predoctores categóricos. Representamos por $\pi\left(X_1,\cdots,X_p\right)$ la

probabilidad de que la característica analizada esté presente condicionada al conjunto de variables $X_1,\cdots,X_p$; esto es:

$$\pi\left(X_1,\cdots,X_p\right)=\mathbf{P}\left(Y=1\middle|X_1,\cdots,X_p\right)$$

El modelo logístico es un tipo de modelo lineal generalizado donde la función de linkage es la función *logit*. Más concretamente:

$$\text{logit}\left(\pi\left(X_1,\cdots,X_p\right)\right)=\beta_0+\beta_1\cdot X_1+\cdots+\beta_p\cdot X_p$$

donde $\text{logit}(x)=\log\left(x/(1-x)\right)$. De esta forma se tiene:

(2.9.1.) $$\pi\left(X_1,\cdots,X_p\right)=\frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1+\exp(\boldsymbol{\beta}'\mathbf{x})}$$

donde $\exp(\boldsymbol{\beta}'\mathbf{x})=\beta_0+\beta_1\cdot X_1+\cdots+\beta_p\cdot X_p$. Esta modelación garantiza que $\pi\left(X_1,\cdots,X_p\right)$ es una cantidad comprendida entre 0 y 1.

Tal como se ha indicado anteriormente, los predictores que aparecen en el modelo deben obviamente ser numéricos. No obstante, cuando se desee introducir un predictor categórico, éste puede desdoblarse a través de un conjunto de variables dummies. Así por ejemplo, supóngase una variable categórica X con tres valores que representaremos por x_1,x_2,x_3. Esta variable puede determinarse de varias formas a través de dos variables dummies D_1 y D_2. Un posible diseño es:

X	D_1	D_2
x_1	**1**	**0**
x_2	**0**	**1**
x_3	**0**	**0**

La forma del diseño generalmente dependerá de los objetivos del investigador. En general, una variable categórica X con k valores puede desdoblarse en k-1 variables dummies a través del diseño que se considere oportuno. De esta forma, en la expresión del modelo logístico, las variables categóricas se introducen a través de sus correspondientes variables dummies.

2.9.2. Significado de los coeficientes del modelo: odd-ratio ajustada.

El objetivo esencial de los modelos de regresión logística en los estudios de caso-control es el cálculo de las odd-ratios ajustadas. Como veremos posteriormente, en el contexto de este tipo de estudio, no es posible en general hacer predicciones de la variable Y a través de un conjunto de variables predictoras. Consideremos por tanto el conjunto de variables predictoras $X_1,\cdots,X_j,\cdots,X_p$. Para la variable X_j analizaremos el efecto de pasar de un valor x a un valor x+1. Se tiene entonces:

$$\text{logit}\left(\pi\left(X_1,\cdots,X_j=x+1,\cdots,X_p\right)\right)=\beta_0+\beta_1\cdot X_1+\cdots+\beta_j\cdot(x+1)+\cdots+\beta_p\cdot X_p$$

análogamente:

$$\text{logit}\left(\pi\left(X_1,\cdots,X_j=x,\cdots,X_p\right)\right)=\beta_0+\beta_1\cdot X_1+\cdots+\beta_j\cdot x+\cdots+\beta_p\cdot X_p$$

La diferencia entre la primera y segunda expresión da lugar a:

$$\log\left(\frac{\pi\left(X_1,\cdots,X_j=x+1,\cdots,X_p\right)\left(1-\pi\left(X_1,\cdots,X_j=x,\cdots,X_p\right)\right)}{\pi\left(X_1,\cdots,X_j=x,\cdots,X_p\right)\left(1-\pi\left(X_1,\cdots,X_j=x+1,\cdots,X_p\right)\right)}\right)=\beta_j \qquad (2.9.2.)$$

Nótese que el argumento del logaritmo en el primer miembro corresponde, para un conjunto de valores fijos de las variables $X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_p$, a la odd-ratio que mide la asociación entre la variable de respuesta Y, y los grupos determinados por $X_j=x$ y $X_j=x+1$. Particularmente, si X_j es una variable binaria con valores 1 y 0

representando la presencia y ausencia respectivamente de una determinada característica, el referido argumento es la odd-ratio correspondiente a la asociación entre las variables binarias Y y X_j. De esta forma, $\exp(\beta_j)$ es propiamente la odd-ratio correspondiente a la asociación entre el predictor X_j y la variable de respuesta Y, ajustada por el resto de variables $X_1,\cdots,X_{j-1},X_{j+1},\cdots,X_p$. Obviamente, la no asociación se presenta cuando $\beta_j = 0$, o lo que es equivalente, cuando la odd-ratio $\exp(\beta_j) = 1$.

2.9.3. Estimación del modelo logístico: método de la máxima verosimilitud.

Insistimos en que el desarrollo del modelo de regresión logística se está realizando en el contexto de un estudio en el que, condicionalmente a un conjunto de predictores $\mathbf{X} = (X_1,\cdots,X_p)$, se observa una variable binaria (aleatoria) Y, la cual indica la presencia o ausencia de algún carácter de interés. En este contexto se presenta la estimación por el método de máxima verosimilitud del modelo, a partir del conjunto de datos de la forma $\{(\mathbf{x}_i, y_i) : i = 1,\cdots,n\}$. De esta forma, la contribución de la observación $(\mathbf{x}_i, y_i)$ a la verosimilitud tiene la forma $\pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i}$, lo que supone que la log-verosimilitud puede expresarse como:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{ \boldsymbol{\beta}'\mathbf{x}_i y_i - \log\left(1+\exp(\boldsymbol{\beta}'\mathbf{x}_i)\right)\right\}$$

El estimador de máxima verosimilitud para el parámetro $\boldsymbol{\beta}$, supuesto que exista, se obtendrá resolviendo el sistema no lineal de p+1 ecuaciones con incógnitas $\boldsymbol{\beta}' = (\beta_0,\beta_1,\cdots,\beta_p)$ de la forma:

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \mathbf{0}$$

Albert y Anderson (1984) dan condiciones bajo las cuales existe el estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}$ en el contexto general de variables con respuesta categórica.

Tales condiciones pueden darse de forma más simple para modelos con respuesta binaria. La clave de la existencia de $\hat{\boldsymbol{\beta}}$ radica en que haya un cierto solapamiento entre dos subconjuntos de datos. Más concretamente, sean los subconjuntos de $\mathbf{R}^p$, $\{\mathbf{x}_j \in \mathbf{R}^p : y_j = 1\}$ y $\{\mathbf{x}_j \in \mathbf{R}^p : y_j = 0\}$. Si existe un hiperplano que los separa, entonces el estimador de máxima verosimilitud no existe. Cuando hay un determinado solapamiento entre ellos, el estimador de máxima verosimilitud existe y es único. Damos ahora los conceptos de separación completa, cuasi-separación y solapamiento.

Separación Completa. Decimos que los conjuntos $C_1, C_2 \subset \mathbf{R}^p$ están separados completamente si existe $\boldsymbol{\alpha} \in \mathbf{R}^p$, tal que para cualquier $\mathbf{x} \in C_1$, $\boldsymbol{\alpha}' \cdot \mathbf{x} > 0$ y para cualquier $\mathbf{x} \in C_2$, $\boldsymbol{\alpha}' \cdot \mathbf{x} < 0$ ($\boldsymbol{\alpha}' \cdot \mathbf{x} = 0$ es la ecuación de un hiperplano que separa a los conjuntos C_1 y C_2).

Separación cuasi-completa. Decimos que los conjuntos $C_1, C_2 \subset \mathbf{R}^p$ están separados cuasi-completamente si existe $\boldsymbol{\alpha} \in \mathbf{R}^p$, tal que para cualquier $\mathbf{x} \in C_1$, $\boldsymbol{\alpha}' \cdot \mathbf{x} \geq 0$ y para cualquier $\mathbf{x} \in C_2$, $\boldsymbol{\alpha}' \cdot \mathbf{x} \leq 0$, dándose la igualdad en al menos un punto.

Solapamiento. Decimos que los conjuntos $C_1, C_2 \subset \mathbf{R}^p$ están solapados si no presentan ninguna de las situaciones anteriores.

Consideremos ahora los conjuntos de $\mathbf{R}^p$ $C_1 = \{\mathbf{x}_j \in \mathbf{R}^p : y_j = 1\}$ y $C_0 = \{\mathbf{x}_j \in \mathbf{R}^p : y_j = 0\}$. Los siguientes teoremas relacionan la posición relativa de estos conjuntos con la existencia del estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}$.

Teorema 1. Si C_1 y C_0 están completamente separados entonces no existe el estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}$.

Teorema 2. Si C_1 y C_0 están solapados entonces existe el estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}$ y es único.

La existencia de estimador de máxima verosimilitud es algo más compleja en situación de cuasi-separación y no la recogemos en esta memoria.

2.9.4. Estrategias para la construcción de modelos de regresión logística.

2.9.4.1. Construcción del modelo mediante selección prospectiva de variables.

En general, la construcción de modelos para describir la asociación de cualquier variable Y con un conjunto de variables $X_1,\cdots,X_p$ debe llevarse a cabo de forma secuencial. Para los fines de esta memoria proponemos un método prospectivo de selección de variables para el modelo de regresión logística. Tal método se basa en los resultados asintóticos para el estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}$ del parámetro $\boldsymbol{\beta}$. Si $\boldsymbol{\beta}$ es k-dimensional, el estadístico:

$$-2\left(l(\boldsymbol{\beta})-l\left(\hat{\boldsymbol{\beta}}\right)\right)\approx\chi^2(p) \tag{2.9.3.}$$

donde por $\chi^2(p)$ entendemos la distribución de probabilidad *ji-cuadrado* centrada y con p grados de libertad. Considérese por tanto un modelo que incluya las variables $X_1,\cdots,X_{r-1}$ y que por tanto, el parámetro $\boldsymbol{\beta}$ sea el vector r-dimensional $\boldsymbol{\beta}_{(r)}=(\beta_0,\beta_1,\cdots,\beta_{r-1})$. En tal caso, $-2\left(l\left(\boldsymbol{\beta}_{(r)}\right)-l\left(\hat{\boldsymbol{\beta}}_{(r)}\right)\right)\approx\chi^2(r)$. Para el modelo que incluya, además de las variables anteriores, la variable X_r, el parámetro $\boldsymbol{\beta}$ correspondiente al modelo de regresión logística tendrá la forma $\boldsymbol{\beta}_{(r+1)}=(\beta_0,\beta_1,\cdots,\beta_{r-1},\beta_r)$. Bajo la hipótesis nula $H_{r,0}:\beta_r=0$, es inmediato que:

i. $l\left(\boldsymbol{\beta}_{(r)}\right)=l\left(\boldsymbol{\beta}_{(r+1)}\right)$

ii. $-2\left(l\left(\boldsymbol{\beta}_{(r+1)}\right)-l\left(\hat{\boldsymbol{\beta}}_{(r+1)}\right)\right)\approx\chi^2(r+1)$

De todo lo anterior puede deducirse que, bajo la referida hipótesis $H_{r,0}:\beta_r=0$, el test estadístico $G_i=2\left(l\left(\hat{\boldsymbol{\beta}}_{(r+1)}\right)-l\left(\hat{\boldsymbol{\beta}}_{(r)}\right)\right)\approx\chi^2(1)$. Este resultado asintótico puede obtenerse a partir del desarrollo de la log-verosimilitud $l(\boldsymbol{\beta})$ en un entorno del

estimador de máxima verosimilitud $\hat{\boldsymbol{\beta}}$. Para un análisis más detallado, ver Cox y Hinkley (1974). Este resultado es suficiente para justificar el siguiente algoritmo de selección de variables para la construcción de un modelo de regresión logística.

Para el conjunto de datos $\{(\mathbf{x}_i, y_i) : i = 1, \cdots, n\}$ donde $\mathbf{x}_i$ es la i-ésima observación del vector de variables continuas $\mathbf{X} = (X_1, \cdots, X_p)$, consideramos el siguiente algoritmo de inclusión prospectiva de variables.

Paso 1. Se estima inicialmente un modelo que incluye exclusivamente el término independiente. De la forma de la verosimilitud, es fácil comprobar que $\hat{\beta}_0 = \text{logit}(\bar{y})$

Paso 2. Para cada variable X_i consideramos el modelo de regresión logística $\text{logit}(\pi(X_i)) = \beta_0 + \beta_i \cdot X_i$ y la hipótesis nula $H_{i,0} : \beta_i = 0$. Bajo esta hipótesis, el test estadístico $G_i^{(1)} = 2\left(l(\hat{\boldsymbol{\beta}}_{1,i}) - l(\hat{\beta}_0)\right) \approx \chi^2(1)$, siendo $\hat{\boldsymbol{\beta}}_{1,i}$ el estimador de máxima verosimilitud del parámetro $\boldsymbol{\beta}'_{1,i} = (\beta_0, \beta_i)$. Para un nivel de significación elegido α, se introduce en el modelo aquella variable X_i para la cual el test estadístico $G_i^{(1)}$ sea máximo, siempre y cuando $G_i^{(1)} \geq \chi_\alpha^2(1)$ donde $\chi_\alpha^2(1)$ es el cuantil 1-α de la distribución $\chi^2(1)$. Si ninguna variable alcanza la significación ($G_i^{(1)} \geq \chi_\alpha^2(1)$), el procedimiento se detiene y no se introduce ninguna variable en el modelo.

Paso 3. Supuesto que en el modelo se hayan introducido las variables $X_1, \cdots, X_r$, para todas aquellas variables X_i que no pertenezcan al modelo, se consideran los modelos de regresión logística de la forma:

$$\text{logit}(\pi(X_1, \cdots, X_r, X_i)) = \beta_0 + \beta_1 \cdot X_1 + \cdots, \beta_r \cdot X_r + \beta_i \cdot X_i$$

y la hipótesis nula $H_{i,0} : \beta_i = 0$. Bajo esta hipótesis, el test estadístico $G_i^{(r)} = 2\left(l(\hat{\boldsymbol{\beta}}_{r,i}) - l(\hat{\boldsymbol{\beta}}_r)\right) \approx \chi^2(1)$, siendo $\hat{\boldsymbol{\beta}}_{r,i}$ el estimador de máxima verosimilitud del

parámetro $\boldsymbol{\beta}'_{r,i} = (\beta_0, \beta_1, \cdots, \beta_r, \beta_i)$ y $\hat{\boldsymbol{\beta}}_r$ el estimador de $\boldsymbol{\beta}'_r = (\beta_0, \beta_1, \cdots, \beta_r)$. Entre todas aquellas variables X_i que alcancen significación $(G_i^{(r)} \geq \chi_\alpha^2(1))$, se introduce en el modelo aquella para la que el test estadístico $G_i^{(r)}$ sea máximo. Este proceso se continúa hasta que no queden variables fuera del modelo que alcancen significación estadística.

2.9.4.2. Comparación de modelos: criterio de información de Akaike.

De acuerdo con lo expuesto en el epígrafe anterior, el estadístico $G_i^{(r)} = 2\left(l\left(\hat{\boldsymbol{\beta}}_{r,i}\right) - l\left(\hat{\boldsymbol{\beta}}_r\right)\right) \geq 0$. La interpretación de este resultado es evidente, a saber: si a un modelo se le añade una nueva variable, la verosimilitud se incrementa. Otra cuestión diferente es que tal incremento de la verosimilitud suponga que la variable añadida tenga significación estadística, tal como se ha visto. Podemos por tanto considerar que $-2 \cdot l\left(\hat{\boldsymbol{\beta}}\right)$ es una medida de desajuste de los datos al modelo, y por tanto, que cuanto menor es $-2 \cdot l\left(\hat{\boldsymbol{\beta}}\right)$, mejor es el modelo. Como se ha puesto de manifiesto, a medida que se introducen nuevas variables en el modelo, disminuye esta cantidad. Ahora bien, un modelo sobredimensionado de variables explicativas no es desde luego la mejor manera de explicar una variable de respuesta *Y*. En este sentido, el criterio de Akaike (ACI) tiene como objetivo medir el ajuste del modelo a partir de la cantidad $-2 \cdot l\left(\hat{\boldsymbol{\beta}}\right)$ penalizada por el número de variables incluidas en dicho modelo. Se define pues la medida de información de Akaike para modelos de regresión logística con respuesta binaria por:

$$\text{AIC} = -2 \cdot l\left(\hat{\boldsymbol{\beta}}\right) + 2 \cdot (p+2) \tag{2.9.4}$$

Esta medida permite comparar dos modelos de regresión logística con respuesta binaria en la forma siguiente: entre dos modelos, es preferible aquél que tiene asociada una menor medida AIC.

2.9.5. Evaluación del ajuste del modelo logístico.

2.9.5.1. Contraste ji-cuadrado de Pearson.

Una vez estimado el modelo, procede evaluar la bondad de ajuste del conjunto de datos $\{(\mathbf{x}_j, y_j): j=1,\cdots,n\}$ al modelo logístico. Para cada $\mathbf{x}_i$, la predicción de $\pi(\mathrm{x}_i)$ tiene la forma:

$$\hat{\pi}(\mathbf{x}_j) = \exp(\hat{\boldsymbol{\beta}}' \cdot \mathbf{x}_j) / \{1 + \exp(\hat{\boldsymbol{\beta}}' \cdot \mathbf{x}_j)\}$$

Para la observación $(\mathbf{x}_j, y_j)$ el residual de Pearson se define por:

(2.9.5.) $$r(y_j, \hat{\pi}(\mathbf{x}_j)) = \frac{y_j - \hat{\pi}(\mathbf{x}_j)}{\sqrt{\hat{\pi}(\mathbf{x}_j) \cdot (1 - \hat{\pi}(\mathbf{x}_j))}}$$

Finalmente, el estadístico ji-cuadrado se define por:

(2.9.6.) $$X^2 = \sum_j r(y_j, \hat{\pi}(\mathbf{x}_j))^2$$

Este estadístico, bajo la hipótesis de que el modelo ajusta adecuadamente a los datos, sigue asíntoticamente una distribución de probabilidad $\chi^2(J-p+1)$. Donde p es el número de covariables (dimensión del vector $\mathbf{X} = (X_1, \cdots, X_p)$) y J *es* el número de clases o patrones, en referencia a cada una de las combinaciones reales de valores del vector $\mathbf{X} = (X_1, \cdots, X_p)$, evidentemente si entre las p covariables hay alguna continua, se tiene $J \approx p$.

2.9.5.2. Estadístico Deviance

El concepto de modelo saturado es útil para valorar el ajuste del modelo al conjunto de datos. La idea es bien simple. Para todos aquellos casos que compartan el mismo vector de variables $\mathbf{x}_j$, se les asigna una probabilidad específica p_j de respuesta $Y_j = 1$. En lo sucesivo, al valor $\mathbf{x}_j$ le llamaremos patrón. Sea por tanto n_j el número de

casos para los que $\mathbf{X}=\mathbf{x}_j$ y m_j el número de éstos para los que $Y_j=1$. La contribución de este conjunto de casos a la verosimilitud viene dada por $p_j^{m_j}\left(1-p_j\right)^{n_j-m_j}$, por lo que el estimador de máxima verosimilitud para p_j tiene obviamente la forma $\hat{p}_j=m_j/n_j$. Si existen J vectores $\mathbf{x}_j$ diferentes, la forma de la log-verosimilitud correspondiente al modelo saturado tiene la forma:

$$l(\text{sat})=l\left(\hat{p}_1,\cdots,\hat{p}_J\right)=\sum_{j=1}^{J}\left\{m_j\cdot\log\left(\frac{m_j}{n_j}\right)+\left(n_j-m_j\right)\cdot\log\left(\frac{n_j-m_j}{n_j}\right)\right\}$$

Supóngase ahora que el modelo considerado es $\text{logit}\left(\pi(\mathbf{x})\right)=\boldsymbol{\beta}'\cdot\mathbf{x}$, siendo nuevamente $\boldsymbol{\beta}'=\left(\beta_0,\beta_1,\cdots,\beta_p\right)$. Se define entonces la *deviance* (distancia de los datos al modelo) como:

(2.9.7)
$$D=2\cdot\left(l(\text{sat})-l\left(\hat{\boldsymbol{\beta}}\right)\right)=$$

$$=2\cdot\sum_{j=1}^{J}\left(m_j\cdot\log\left(\frac{m_j}{n_j\cdot\hat{\pi}\left(\mathbf{x}_j\right)}\right)+\left(n_j-m_j\right)\cdot\log\left(\frac{n_j-m_j}{n_j\cdot\left(1-\hat{\pi}\left(\mathbf{x}_j\right)\right)}\right)\right)$$

Nótese que el modelo saturado se ha parametrizado mediante J parámetros y el modelo considerado con p+1. Esto supone, que bajo la hipótesis de que el modelo ajusta adecuadamente a los datos, el estadístico *deviance* es tal que $D\cong\chi^2\left(J-p+1\right)$. Para más detalles acerca de este estadístico, ver McCullag y Nelder (1983).

2.9.5.3. Test de Hosmer y Lemeshow.

Los contrastes basados en los test estadísticos ji-cuadrado de Pearson X^2 y el estadístico deviance D, podrían no garantizar el correspondiente error alpha. El problema radica en el número J de patrones (valores diferentes del vector predictor $\mathbf{x}_j$). Bajo la hipótesis de que los datos ajusten correctamente al modelo, ambos estadísticos tienen asíntoticamente distribución $\chi^2\left(J-p+1\right)$. Ahora bien, si el valor J crece al mismo ritmo que número de observaciones n, la convergencia no está garantizada. Para

ilustrar esta idea, supóngase que los datos se agrupan en una tabla de contingencia $J\times 2$, donde las filas corresponden a los diferentes valores del predictor $\mathbf{x}_j$, y las columnas a los valores 0 ó 1 de la respuesta y_j. Si $J \approx n$, esto es, si el número de filas aumenta a la misma tasa que el número de datos, obviamente no puede concluirse que la distribución asíntotica para X^2 sea $\chi^2(J-p+1)$. Para obviar este problema, Hosmer y Lemeshow (1980 y 1982) propusieron un *test de ajuste* basado en agrupamientos de los valores de $\mathbf{x}_j$, el cual describimos seguidamente.

Para $J \approx n$, las n predicciones $\hat{\pi}(\mathbf{x}_j)$, se ordenan en n filas de menor a mayor, agrupándose en g clases, habitualmente $g=10$. Esta clasificación puede tener como puntos de corte valores fijos de la probabilidad estimada, o bien los percentiles correspondientes a las referidas predicciones $\hat{\pi}(\mathbf{x}_j)$. Cuando los datos se reparten homogéneamente, los dos tipos de partición son equivalentes. En la organización basada en percentiles y para $g=10$, los puntos de corte son deciles y habitualmente se les llama "deciles de riesgo". Consideramos que en el primer decil de riesgo, que corresponde a las probabilidades más pequeñas, hay $n/10$ elementos, los mismos que en el décimo decil. Para la columna de respuesta Y=1, la probabilidad estimada para cada celda es obviamente la suma de las probabilidades estimadas de todos los componentes del grupo. Para la columna de respuesta Y=0, la probabilidad estimada se obtiene de forma análoga. Con cualquiera de los dos tipos de agrupamientos, en esta tabla así formada $g\times 2$ se define el test estadístico:

$$C=\sum_{k=1}^{g}\frac{(O_k-n'_k\bar{\pi}_k)^2}{n'_k\bar{\pi}_k(1-\bar{\pi}_k)} \qquad (2.9.8.)$$

donde n'_k, es el número total de sujetos en el k-ésimo grupo, siendo O_k y $\bar{\pi}_k$ el número de respuestas Y=1 y la correspondiente probabilidad estimada para ese grupo. Ahora, bajo la hipótesis de que el modelo ajusta adecuadamente a los datos, el test estadístico C, tiene asíntoticamente ($n\to\infty$) una distribución $\chi^2(g-2)$.

2.9.5.4. Observaciones influyentes.

Nos ocupamos ahora de detectar aquellas observaciones $(\mathbf{x}_j, y_j)$ que influyen muy marcadamente en el modelo final, esto es, observaciones que si fueran omitidas de la construcción del modelo, darían lugar a fuertes variaciones en los estimadores. En muchas ocasiones, tales observaciones proceden de individuos ajenos a la población de estudio. Es recomendable por tanto identificarlas, y en su caso, no utilizarlas en la construcción del modelo.

Supongamos nuevamente un modelo con p variables predictivas y J patrones $\mathbf{x}_j$. La matriz de diseño $\mathbf{X}$ será por tanto de dimensión es $J \times (p+1)$. Prebigon, propone para el análisis de valores influyentes la matriz $\mathbf{H}$, llamada *matriz sombrero*:

(2.9.9.) $$\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2}$$

donde $\mathbf{V}$ es una matriz diagonal $J \times J$, cuyos elementos son:

$$v_j = n_j(\hat{\pi}(\mathbf{x}_j) \cdot (1 - \hat{\pi}(\mathbf{x}_j))$$

donde n_j es el número de observaciones que comparten el patrón $\mathbf{x}_j$. El j-ésimo elemento de la diagonal de la matriz $\mathbf{H}$, el cual corresponde al patrón $\mathbf{x}_j$, tiene la forma:

$$h_j = n_j\hat{\pi}(\mathbf{x}_j) \cdot (1 - \hat{\pi}(\mathbf{x}_j))(1, \mathbf{x}_j)(X'VX)^{-1}(1, \mathbf{x}_j)' = v_j x b_j$$

donde $b_j = (1, \mathbf{x}_j)(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}(1, \mathbf{x}_j)'$ (2.9.10.)

Al valor h_j se le suele llamar *influencia* y mide de algún modo la distancia de la observación $(\mathbf{x}_j, y_j)$ del núcleo de datos. Prebigon (1981) demuestra que cuanto mayor el valor de (2.9.10.), para un determinado patrón $\mathbf{x}_j$, mayor es su influencia. La influencia tiene un límite, concretamente en valores con probabilidades estimadas $\hat{\pi}(\mathbf{x}_j) < 0.1$ ó $\hat{\pi}(\mathbf{x}_j) > 0.9$ disminuye asíntoticamente.

A partir del valor residual de Pearson $r\left(y_j,\hat{\pi}\left(\mathbf{x}_j\right)\right)$ (2.9.5.), se puede define el residual estandarizado:

(2.9.11.) $$r_{s,j}=r\left(y_j,\hat{\pi}\left(\mathbf{x}_j\right)\right)\Big/\sqrt{1-h_j}$$

Las observaciones más influyentes tendrán valores h_j altos que se acercarán a uno y por tanto, *inflarán* el valor del residual estandarizado en relación al ordinario.

En esta última fase de análisis del modelo se pueden detectar valores extremos (outliers), que tengan grandes residuales y/o gran influencia. En el caso de prescindir de alguno de estos valores, los estadísticos X^2 y D se reducirán. La correspondiente variación viene dada por:

(2.9.12.) $$\Delta X_j^{\,2}=\frac{r\left(y_j,\hat{\pi}\left(\mathbf{x}_j\right)\right)^2}{1-h_j}=r_{s,j}^2$$

y $$\Delta D_j=\frac{d_j^2}{1-h_j}$$

Estas medidas de análisis individualizado, tienen una expresión gráfica, que resulta de utilidad. En este sentido, las representaciones de diagrama de dispersión de los pares $\left(\hat{\pi},\Delta X^2\right)$ y $\left(\hat{\pi},\Delta D\right)$ pueden detectar valores de gran influencia.

2.9.6. Corrección de la sobreestimación del vector de coeficientes de RL.

En lo que se ha expuesto sobre los modelos de regresión logística, siempre consideramos, implícitamente, un importante volumen inicial de datos, en el que nos apoyábamos para construir el modelo. Esto no siempre es así, y en ocasiones el número de predictores es grande con respecto al número de datos, vulnerando la regla 1-10 (al menos 10 datos por cada variable predictora). Esta situación de escasez de datos, conlleva una *inflación* del vector $\boldsymbol{\beta}$ de coeficientes, sobredimensionándose los valores de las componentes del mismo, lo que repercute sobre las odd-ratios obteniéndose valores totalmente hinchados. En esta misma situación de escasez de datos, también puede afectar al número de variables que finalmente entran a formar parte del modelo.

Dicho más formalmente, cuando la dimensión del vector $\boldsymbol{\beta}$ es *grande* en relación al tamaño muestral *n*, el sesgo del estimador de máxima verosimilitud $b\left(\hat{\boldsymbol{\beta}}\right) = E\left[\hat{\boldsymbol{\beta}}\right] - \boldsymbol{\beta}$ puede ser muy apreciable. Nos proponemos estudiar la última situación arriba descrita: los efectos sobre $\boldsymbol{\beta}$ de la escasez de datos ,su sobreestimación y un método para buscar un factor de corrección que, en estas condiciones, transforme el vector beta en otro más aproximado al que se obtendría si las condiciones fueran ideales.

Previamente el problema de corrección de errores en RL ha sido tratado por Efron B. (1983) , Copas JB (1983), Steyerber E. W. y otros(2000). La herramienta que debe emplearse para estudiar y solucionar la inflación de las componentes de $\boldsymbol{\beta}$ y todo lo que conlleva la escasez de datos, no puede ser otra que la computación. La metodología que se utilizará es el bootstrap, que ya ha sido expuesta en el epígrafe 2.5. Finalmente se darán algoritmos para resolver la cuestión planteada.

2.9.6.1. Corrección Bootstrap.

Se expone en este epígrafe una corrección genérica del sesgo, basada en metodología bootstrap (por ello se utiliza la notación del punto 2.5), que ilustra y sirve de base a posteriores algoritmos. Sea $y = \left(y_1, y_2, \cdots, y_n\right)$ una muestra aleatoria de una distribución de probabilidad F desconocida. Queremos, usando el estimador $\hat{\theta} = s(y)$, estimar el parámetro $\theta = t(F)$.

Estimamos una función de distribución empírica $\hat{F}$, basada en la probabilidad; para cada muestra bootstrap $y^* = \left(y_1^*, \cdots, y_n^*\right)$, estimamos la probabilidad de aparición del *j*-ésimo elemento: $\mathbf{P}_j^* = \left\{y_i^* = y_j\right\}/n$, $1 \le j \le n$. De esta manera definimos el vector probabilístico $\mathbf{P}^* = \left(\mathbf{P}_1^*, \mathbf{P}_2^*, \cdots, \mathbf{P}_n^*\right)$. Ahora las réplicas bootstrap de $\hat{\theta}^* = s(y^*)$ se calculan mediante el vector P^* aplicado a los datos originales, y la estimación $\hat{\theta} = t(\hat{F})$, viene dada por la expresión $\hat{\theta}^* = T(\mathrm{P}^*)$, indicando que la estimación bootstrap de θ, está basada en el vector P^*. Sobre la muestra original, de tamaño *n*, cada elemento tiene la misma probabilidad, $1/n$ de ser elegido. Esta estimación del vector, donde todas las componentes tienen el mismo valor, se designa por P^0; que es

la usada en la estimación de $\hat{\theta}$. De este modo tenemos $\hat{\theta} = t(\hat{F}) = T(\mathbf{P}^0)$. Formamos muestras $y^{*1}, y^{*2}, \ldots\ldots, y^{*b}$, de cada una de ellas estimamos su vector probabilístico : $\mathrm{P}^{*1}, \mathrm{P}^{*2}, \cdots, \mathrm{P}^{*B}$. Definimos $\overline{\mathbf{P}}^*$, como la media de los vectores probabilísticos correspondientes a cada muestra:

$$\overline{\mathbf{P}}^* = \sum_{b=1}^{B} \mathbf{P}^{*b} \Big/ B$$

Basándonos en esta estimación de probabilidad la corrección del sesgo puede venir dada por $\hat{b}_B = \hat{\theta}^*(.) - T(\mathrm{P}^0)$. Esta expresión puede ser mejorado por otra, basada en el vector probabilístico, dada por $\overline{b}_B = \hat{\theta}^*(.) - T(\overline{\mathrm{P}}^*)$, donde todos los elementos se hallan definidos en este epígrafe.

Cabe preguntarse qué diferencia existe entre $\hat{b}_B$ y $\overline{b}_B$: el primero $\hat{b}_B$ converge al bootstrap ideal $\hat{b}_\infty$, propiedad que también cumple $\overline{b}_B$, pero en este último caso la rapidez de convergencia es mayor.

2.9.6.2. Método del Factor de Corrección Lineal.

Existen varias alternativas para solucionar el problema de la inflación del vector $\boldsymbol{\beta}$ de RL. La primera es el método de la *Máxima Verosimilitud Penalizado* (Hoerl and Kennard 1970) y aplicado a regresión logística por Van Houwelingen (1992). El segundo método es el *Lasso (Least Absolute Shrinkage and Selection Operator)* se debe a Tibsharani (1996). La última se conoce como *Factor de Corrección Lineal*; es la desarrollada en esta memoria. Figura en un artículo publicado por E. W. Steyerberg, y otros, en la revista *Statistics in Medicine* (2000) e ilustrada con un estudio de casos en *Statistica Neerlandica (2001)*. Se debe principalmente a F. E. Harrell.

Esta última propuesta trata de , encontrándonos en estas condiciones de escasez de datos, buscar un factor lineal de corrección *(shrinkage factor)* tal que aplicado al vector beta sobreestimado, elimine esta inflación; considerando así un nuevo vector, cuyas componentes tomen valores más próximos a los obtenidos con un gran volumen de datos. Este método también se conoce como *corrección después de encajar el modelo.*

Partimos de un conjunto de datos $\{(\mathbf{x}_i, y_i) : i = 1, \cdots, n\}$. La idea central subyacente es reducir los p predictores de un individuo de la base de datos, conjuntamente con el vector beta, a un solo valor, llamado *Índice Pronostico (IP)*; con ello habremos reducido un problema p-dimensional a otro unidimensional, modelo este donde no existe este problema de sobreestimación. Consideramos un vector de variables predictoras $X_1, \cdots, X_p$, o, lo que es lo mismo, un vector p-dimensional y una variable respuesta Y bidimensional. Sea $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$ el vector beta del modelo de RL en los términos expresados en el epígrafe 2.9.1. Definimos el *índice Pronóstico* para el i-ésimo individuo como:

(2.9.13.) $$IP_i = \sum_{j=1}^{p} \beta_j x_{j_i} \ ; \text{ donde } i = 1, \cdots, n;$$

Nótese que, con esta definición, para cada individuo i, $(i = 1, \cdots, n)$, la variable respuesta, Y_i, no se ve alterada, además se ha obtenido un par (Y_i, IP_i) para cada individuo.

Ajustamos un nuevo modelo de RL en este caso unidimensional, de la siguiente manera:

(2.9.14.) $$\text{logit}\{\mathbf{P}(Y = 1 | IP_i)\} = \gamma_i \cdot IP_i \ : \ i = 1, \cdots, n$$

Esta es una de las cuestiones clave : al ser la RL unidimensional, el efecto inflación no se produce. Evidentemente cada γ_i ; $i = 1, \cdots, n$; es un factor de corrección

2.9.6.3. Algoritmo de Harrell.

Este algoritmo consiste en estimar, por generación bootstrap, vectores beta ($\boldsymbol{\beta}^*$), y, aplicándolos a los datos originales, construir índices pronostico (*IP*). Ajustar el modelo de RL unidimensional como en (2.9.14), calculando de este modo un *Factor de Corrección*, γ_i, para cada sujeto, la media de los factores, para un vector beta dado, es el *FC* para esa generación bootstrap. Se reitera el proceso para cada vector $\boldsymbol{\beta}^*$ bootstrap . El *FC* final es la media de todos los factores calculados de este modo.

Partimos de una muestra de individuos $\{(\mathbf{x}_i, y_i) : i = 1,\cdots,n\}$ (datos originales), donde cada $\mathbf{x}_i$, $i = 1,\cdots,n$ es un vector predictor p-dimensional, con una respuesta binaria Y_i, $i = 1,\cdots,n$.

Paso 1.- Seleccionamos, con reemplazamiento, de la muestra original $\{(\mathbf{x}_i, y_i) : i = 1,\cdots,n\}$, otra aleatoria de tamaño *n*: $\mathbf{x}_1^*,\cdots,\mathbf{x}_n^*$, muestra bootstrap.

Paso 2.- Con la muestra seleccionada $\mathbf{x}_1^*,\cdots,\mathbf{x}_n^*$, estimamos el modelo de RL, calculando el vector $\boldsymbol{\beta}_{(i)}^*$. Esto es, generación bootstrap de datos y consecuentemente de vector beta asociado.
Entendemos que la estimación del modelo se ha realizado incluyendo una selección de variables predictoras.

Paso 3.- Con el $\boldsymbol{\beta}_{(i)}^*$ seleccionado, calculamos los índices pronóstico (*IP*) para todos los sujetos de la muestra original. Se aplica la expresión (2.9.13.).

Paso 4.- Con los valores originales de la variable outcome *Y,* ajustamos la (2.9.14.). Una vez obtenidos los *n* factores de corrección puntuales FC's $(\gamma^j{}_i)$, donde *i* representa la generación bootstrap y *j* el individuo, calculamos la media de los mismos:

$$\gamma_i = \frac{\sum_{j=1}^{n} \gamma_j^{(i)}}{n}$$

Paso 5.- Los *pasos 1,2,3 y 4* se repiten *B* veces, con lo que se obtienen $\gamma^{(i)}$; $i = 1,\cdots,B$; factores de corrección.

Paso 6.- El factor de corrección final es la media de los FC's calculados en el *Paso 5.*

(2.9.15.)
$$FC = \gamma = \frac{\sum_{i=1}^{B} \gamma^{(i)}}{B}$$

Se recomienda tomar *B=10*n.* A este respecto consultara epígrafe 2.5.

2.9.6.4. Corrección al algoritmo de Harrell.

El algoritmo expuesto en el punto anterior, identificado como de *Harrell puro,* obtiene un factor de corrección (*shrinkage factor*), γ ,para el vector de coeficientes. Sin embargo, empíricamente se demuestra que este factor corrector es muy fuerte y aplicado al vector beta de coeficientes de RL, no solo elimina la inflación del mismo, sino que incluso el valor de las componentes del vector corregido están por muy por debajo del valor obtenido para beta con un gran número de datos (véase epígrafe 2.9.6.5). Considerando este inconveniente, y con el fin de obtener un vector más próximo al que obtendríamos sin escasez de datos, proponemos un nuevo factor de corrección γ_c de *Harrell corregido*, consistente en eliminar los valores más pequeños del factor de corrección obtenidos por generación bootstrap. De este modo el valor de γ_c será ligeramente superior al de γ , con lo que el nuevo vector estará más próximo al estimado en condiciones óptimas. Se realiza una eliminación de factores mínimos muy conservadora, pues tan solo un 1% son eliminados. . Para ilustrar este punto se ha realizado una simulación (véase 2.9.6.5.).

En algoritmo para el cálculo de γ_c, es muy parecido al contenido en el epígrafe 2.9.6.3. Los *Pasos 1 a 5* son análogos, a partir de este continúa de la siguiente manera:

Paso 6.- De los *B* factores calculados por generación bootstrap localizar y eliminar (en el sentido que no entran en la estimación de la media) el 1% de valores más pequeños.

Paso 7.- El factor de corrección final es la media de los FC's calculados en el punto 6.

(2.9.16.)
$$FC = \gamma_c = \frac{\sum_{i=1}^{0.99*B} \gamma^{(i)}}{0.99*B}$$

El número de generaciones bootstrap no se altera respecto al algoritmo anterior.

2.9.6.5. Panorámica de resultados: Simulación.

Con la doble finalidad de dar una amplia panorámica de cómo, según se reduce el numero de casos, existen variables que no entran en el modelo y sobre la inflación del vector beta de coeficientes; y por otro lado, examinar en la práctica el funcionamiento de los factores de corrección, tanto de modo general, como comparativamente entre los dos; se ha realizado una simulación con 10 variable predictoras independientes, que para n=988, están todas en el modelo. Se va reduciendo el número de datos, y a diversas cantidades de datos, se analizan las variables que entran en el modelo, sus coeficientes y los resultados obtenidos al aplicar los dos factores de corrección. Las diez variable predictoras: *edad* (continua) y por la que ajustamos todas la demás, cinco variables con distribución normal que representan diferentes variables biológicas (*colesterol, peso, ácido úrico, glucemia, triglicéridos*), que han sido codificadas de acuerdo con Tablas 3.2 y 3.3. Por último cuatro variable binarias, que representan *sexo*, *diabetes*, *tabaco* y otro *factor indeterminado*. La primera variable predictora es un factor de protección. Los resultados obtenidos para n= 988 (resultados reales), n=591, n=351, n=221 y n=78, este último vulnerando la regla 1-10; se exponen en Tabla 2.9.17.

Por construcción del modelo, al ser todas las variables independientes, con un número suficiente de datos, todas están en modelo; así para n=988 , fila en verde, es el modelo real. Para n=591 solo entran p= 9 variables en el modelo (última columna a la derecha), sucesivamente para n=591 se obtiene p=9, n=351 p=8, n=321 p=5 , y n=74 p=2. La inflación de los coeficientes, si bien en las primeras reducciones de datos (n=591 y n=351, dos primeros bloques en fila, después de los coeficientes reales; no se manifiesta claramente, cuando la reducción es importante (n=221 y n=78, dos últimos bloques en fila) se hace evidente; aún mas vulnerando la regla 1-10 (n=78, última fila en verde), ya solo quedan 2 variables en el modelo y sus coeficientes están claramente sobredimensionados.

Especial mención merece β_1, pues su coeficiente siempre es negativo (factor de protección); este coeficiente, en escasez de datos, también se incrementa, se podría interpretar que levemente; pero hemos de recordar que se trata de una variable continua, y ese incrementos se registra por cada unidad aumentada (año transcurrido, si fuera la edad).

Tabla 2.9.17. Simulación vector beta de coeficientes y factores correctores.

Nº Cas.	$\beta_1\,(se)$	$\beta_2\,(se)$	$\beta_3\,(se)$	$\beta_4\,(se)$	$\beta_5\,(se)$	$\beta_6\,(se)$	$\beta_7\,(se)$	$\beta_8\,(se)$	$\beta_9\,(se)$	$\beta_{10}\,(se)$	Vars.
n=988	-0.043 (.004)	.669 (.158)	.368 (.196)	.925 (.139)	.624 (.142)	.657 (.199)	.498 (.130)	.634 (.155)	.778 (.144)	.660 (.144)	p=10
n=591	-0.041 (.005)	.529 (.206)		.869 (.183)	.772 (.187)	1.030 (.294)	.392 (.170)	.598 (.202)	.834 (.190)	.902 (.187)	p=9
FC γ =.404	-0.057()	0.214		.351	.312	.416	.158	.241	.337	.364	
FC γ_c =.425	-0.058	0.225		.369	.328	.438	.167	.254	.354	.383	p=9
n=351	-0.037 (.006)	.808 (.259)		.673 (.233)	.638 (.232)		.550 (.207)	.789 (.254)	.471 (.238)	.669 (.241)	p=8
FC γ =.370	-0.051	.299		.249	.236		.203	.292	.174	.247	p=8
FC γ_c =.373	-0.050	.301		.251	.238		.205	.295	.176	.249	p=8
n=221	-0.034 (.006)	1.085 (.339)		1.166 (.312)	1.070 (.299)			1.127			p=5
FC γ =.255	-0.0426	.277		.297	.273			.288			p=5
FC γ_c =.260	-0.0428	.282		.303	.278			.293			p=5
n=78	-0.014 (.006)			1.873 (.495)							p=2
FC γ =.135	-0.0158			.253							p=2
FC γ_c =.138	-0.0159			.258							p=2

Respecto a los factores de corrección, ocurre lo mismo que con la inflación de los coeficientes: cuando la reducción de datos es moderada, al no manifestarse claramente el sobredimensionamiento, no vale la pena aplicarlos, pues la reducción aplicada a cada β_i, $\forall i = 1,...,10.$ es drástica. Sin embargo cuando la reducción es importante, n= 221 y n=78 , la aplicación de los FC resulta imprescindible y sitúan los coeficientes de las variables, que aún están en la ecuación en valores mucho más cercanos a los reales. A este respecto fijémonos, para n = 78 (última fila de color verde), que vulnera la regla1-10, en los coeficientes de edad y sexo: se encuentran totalmente sobre dimensionados, la aplicación de los FC's los transforma en valores más próximos a los reales. En parecidos términos nos podríamos expresar sobre el bloque correspondiente a $n = 221$.

Comparativamente los resultados obtenidos por los dos FC's , γ y γ_c, se expresan en cada bloque, en dos filas consecutivas de la tabla, basta un somero análisis, para ver que γ_c, obtiene mejores aproximaciones, siempre a la baja sobre el valor real, que γ.

2.9.7. Estudios de caso- control.

El conjunto de datos a analizar $\{(\mathbf{x}_j, y_j): j = 1,\cdots,n\}$ en los estudios de caso-control tiene la misma forma que en los estudios de cohorte. Sin embargo ahora, la variable binaria Y (con valores 0 y 1), es la variable de estratificación, y por tanto no es una variable aleatoria. El diseño del estudio de caso control es tal, que para cada valor de la variable Y, se observa el vector de variables **x**. Ello supone que la verosimilitud es de naturaleza diferente. Sin embargo, en este epígrafe mostraremos cómo, solo a efectos de estimación de odd-ratios, la verosimilitud correspondiente al conjunto de datos es equivalente para ambos tipos de diseño.

La función de verosimilitud correspondiente al diseño de caso-control tiene la forma:

(2.9.18.) $$L(\beta) = \prod_{i=1}^{n_1} \mathbf{P}(x_i \mid y_i = 1) \cdot \prod_{i=1}^{n_0} \mathbf{P}(x_i \mid y_i = 0) \qquad \text{Con} \quad n_0 + n_1 = n;$$

La variable que de interés, independiente es Y, para transformar la expresión anterior de tal manera que la variable Y figure como independiente, tomamos un variable de selección S que índica si un individuo ha resultado elegido en la muestra. Con esta transformación la (2.9.18) toma la forma:

$$L(\beta)=\prod_{i=1}^{n_1}\mathbf{P}(x_i \mid y_i=1,S_i=1)\cdot\prod_{i=1}^{n_0}\mathbf{P}(x \mid y_i=0,S_i=1) \tag{2.9.19}$$

Usando el teorema de Bayes cada factor de la verosimilitud puede expresarse como:

$$\mathbf{P}(x_i \mid Y_i,S_i=1)=\frac{\mathbf{P}(Y_i \mid x_i,S_i=1)\cdot\mathbf{P}(x_i \mid S_i=1)}{\mathbf{P}(Y_i \mid S_i=1)}$$

Volviendo a utilizar el teorema de Bayes, obtenemos:

$$\mathbf{P}(Y=1 \mid x,S=1)=\frac{\mathbf{P}(Y=1 \mid x)\mathbf{P}(S=1 \mid x,Y=1)}{\mathbf{P}(Y=0 \mid x)\mathbf{P}(S=1 \mid x,Y=0)+\mathbf{P}(Y=1 \mid x)\mathbf{P}(S=1 \mid x,Y=1)} \tag{2.9.20.}$$

Asumimos que la selección de los casos y controles se ha hecho independiente de los predictores , con probabilidades τ_1 y τ_0 , obtenemos:

$$\tau_1=\mathbf{P}(S=1 \mid Y=1,x)=\mathbf{P}(S=1 \mid Y=1)$$

del mismo modo:

$$\tau_0=\mathbf{P}(S=1 \mid Y=0,x)=\mathbf{P}(S=1 \mid Y=0)$$

Sustituyendo estas expresiones en (2.4.20.) obtenemos:

$$\mathbf{P}(Y=1 \mid x,S=1)=\frac{\tau_1\pi(x)}{\tau_0[1-\pi(x)]+\tau_1\pi(x)}$$

Operando esta última expresión y simplificando, obtenemos:

$$\mathbf{P}(Y=1 \mid x, S=1)=\frac{\exp(\beta_0^*+\beta_1 x)}{1+\exp(\beta_0^*+\beta_1 x)}$$

donde $\beta_0^*=\beta_0 \log\left(\tau_1/\tau_0\right)$

Sustituyendo en (2.9.20.) y teniendo en cuenta la independencia entre x y S; esta expresión se transforma en:

$$\square(x_i \mid Y=1, S_i=1)=\frac{\dfrac{\exp(\beta_0^*+\beta_1 x)}{1+\exp(\beta_0^*+\beta_1 x)}\cdot \square(x)}{\square(Y=1 \mid S_i=1)}$$

Del mismo modo, para Y=0 se obtiene la expresión:

$$\square(x_i \mid Y=0, S_i=1)=\frac{\dfrac{1}{1+\exp(\beta_0^*+\beta_1 x)}\cdot \square(x)}{\square(Y=0 \mid S_i=1)}$$

A partir de estas dos últimas expresiones, la función de Verosimilitud (2.9.18.) puede expresarse de la siguiente manera:

$$L(\beta)=\prod_{i=1}^{n_1}\left[\frac{\exp(\beta_0^*+\beta_1 x)}{1+\exp(\beta_0^*+\beta_1 x)}\right]^{y_i}\cdot\left[\frac{1}{1+\exp(\beta_0^*+\beta_1 x)}\right]^{1-y_i}\cdot\prod_{i=1}^{n_0}\left[\frac{\mathbf{P}(x)}{\mathbf{P}(y_i \mid S_i=1)}\right]$$

El último factor del producto de la expresión anterior, no depende de β. La expresión de la verosimilitud $L(\beta)$ para un estudio de caso-control es equivalente a la de un estudio de cohortes, si bien la utilidad en caso-control se ciñe a la estimación de odds-ratios.

Capítulo III: Estudio de hipertensión arterial en Lanzarote.

3.1. Introducción.

La isla de Lanzarote, junto a los islotes de La Graciosa y Alegranza, son los territorios más septentrionales del Archipiélago Canario. Tiene una extensión de 845 Km^2 y dista de la costa africana unas 55 millas náuticas. Posee una población de derecho que asciende a unas 118.000 personas, a las que debemos añadir unos 50.000 turistas, fundamentalmente procedentes de centro Europa e Inglaterra, que rotativamente disfrutan aquí sus vacaciones.

La Isla es de suave orografía, lo que no favorece la lluvia, siendo la media pluviométrica mensual de 4.9 mm. La escasez de agua es un problema endémico, subsanado mediante la desalinización de agua de mar. La temperatura media es de 21.5°C. Durante casi todo el año soplan los alisios, que con su humedad favorecen la escasa agricultura. La temperatura tan benigna ha favorecido el desarrollo del turismo, en detrimento de las fuentes económicas tradicionales de agricultura y pesca.

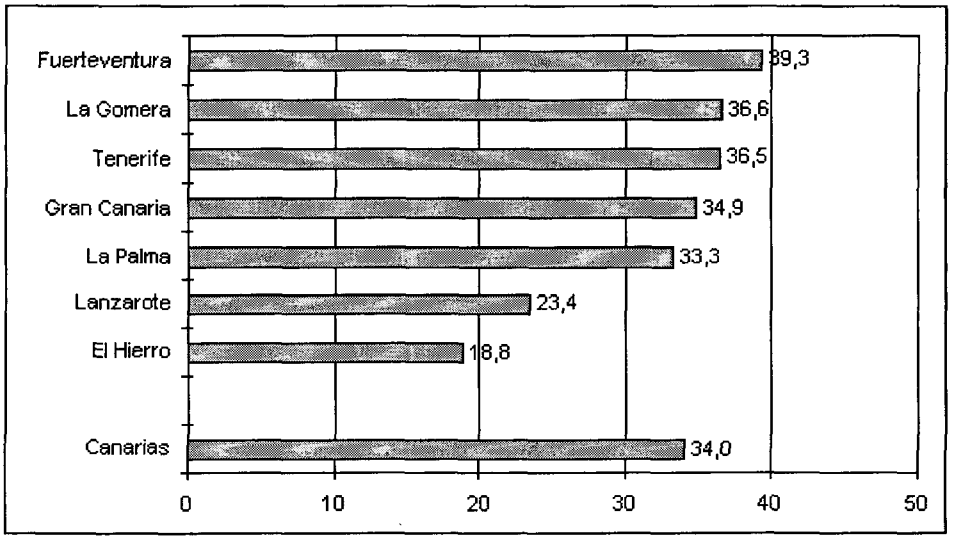
Los habitantes de la Isla presentan diferentes orígenes. Desde etnia bereber, un gran número tiene antepasados de España peninsular, hasta grupos con raíces centroeuropeas, francesas o portuguesas. No existe factor de consanguinidad.

Los hábitos alimenticios han cambiado sustancialmente. Se ha pasado de una dieta basada en pescado y productos frescos del campo, a un tipo de comida congelada, carnes y pescados importados, deficitaria en frutas y verduras. El consumo del gofio (harina de maíz tostado) alimento tradicional, ha disminuido drásticamente.

Lanzarote sanitariamente se divide en siete zonas básicas de salud, y cuenta con tres hospitales. El Hospital General pertenece al Servicio Canario de Salud, cuenta con 190 camas, y en él se practican la mayoría de las especialidades médicas. El Hospital Insular, que pertenece al Cabildo, cuenta con 190 camas y actúa fundamentalmente como centro geriátrico. Ambos están situados en la capital, Arrecife. Por último, existe un hospital privado situado en Puerto del Carmen, que cuenta con 52 camas.

De acuerdo con la encuesta nutricional realizada en Canarias correspondiente al periodo 1997-1998, la prevalencia de hipertensión en Lanzarote es de un 23,4%. La siguiente figura muestra las prevalencias de hipertensión arterial en cada una de las islas del archipiélago.

Tabla 3.1.1.



En este capítulo se examinan algunos factores asociados con la hipertensión arterial (HTA) en Lanzarote así como posibles consecuencias de la misma a partir de los datos aportados por un estudio de caso-control, en el que la variable de estratificación es la indicatriz de presencia de HTA. No nos es por tanto posible dar una estimación de prevalencia de HTA.

El conjunto de datos en los que se ha basado el estudio se describe en el epígrafe 3.2 del presente capítulo. A este conjunto de datos, se les han aplicado los métodos de análisis estadístico para estudios de caso-control revisados en el capítulo anterior. Para tal fin, además de utilizarse el paquete estadístico SPSS, se ha elaborado un software preciso, tanto en lenguaje Pascal 7.0 o C++, como en el lenguaje propio del entorno S/R. Concretamente se han implementado los métodos de estimación bootstrap, un algoritmo para estimación no paramétrica de curvas ROC así como el cálculo del factor de corrección lineal para el vector de regresión logística (véase Apéndice II). El CD-Rom que acompaña a esta memoria contiene los programas ejecutables y la base de datos en soporte SPSS.

En el epígrafe 3.3 se resumen los principales resultados del estudio. No debe olvidarse que la finalidad principal de esta memoria es metodológica, por lo que en la tabla 3.3.7 se discuten diferentes formas de resumen de los datos y métodos de comparación. Las asociaciones crudas y ajustadas por edad de las variables numéricas y categóricas se muestran en las tablas 3.3.7 y 3.3.9. Dada la extensión y complejidad de las tablas 3.3.6 y 3.3.7, por razones didácticas, se discuten conjuntamente con la tabla, aunque las conclusiones se encuentran en el último capítulo. También se realiza el análisis preliminar de factores categóricos, contenido en la tabla 3.3.9.

El epígrafe 3.4 contiene los resultados de la estimación de curvas ROC, obteniéndose comparativamente por tres métodos: estimación cruda, normal y no paramétrica. Se ha aplicado a dos ejemplos, variables *edad* y *glucemia*, como marcadores de hipertensión arterial. El punto 3.5 contiene los resultados del análisis discriminante.

En el epígrafe 3.6 se presentan los resultados obtenidos en la metodología de regresión logística; se presentan modelos para las variables *hta, corazón* y *cerebro*. Con el fin de dar una panorámica sobre esta metodología, en cada uno de los procesos

se estima lo fundamental: el modelo y sus odds-ratios; además de aquellos resultados interesantes para cada estimación del modelo, bien sean residuos, tablas de clasificación o algún tipo de test.

Por último, en el punto 3.6.4 contiene los resultados de la metodología computacional bootstrap, aplicada a corrección de la inflación registrada en el vector beta de coeficientes de RL. Con esta finalidad, se ha confeccionado una herramienta informática (véase Apéndice II).

3.2. Material y población.

Los datos que se tratan en esta memoria han sido tomados de individuos adultos de la zona de salud básica número uno de Lanzarote, que abarca los siguientes barrios de Arrecife: Santa Coloma, Argana Alta, Argana Baja, Maneje, La Vega y San Francisco Javier. El diseño del estudio es de caso-control, habiéndose seleccionado una muestra de 368 personas con hipertensión arterial establecida (casos) y otra de 128 personas normotensas (controles). En el grupo de hipertensos, 162 eran hombres (44%) y 206 mujeres (56%), mientras que en el de normotensos, 46 eran hombres (36%) y 82 mujeres (64%). En cada grupo se han determinado además de la edad, variables antropométricas tales como el peso y la talla y las variables binarias indicatrices de tabaquismo y ejercicio físico. Se han determinado asimismo variables relacionadas con afecciones cardiacas y cerebro-vasculares. La presencia de diabetes se ha determinado a través de las determinaciones de glucemia. Se han observado también las variables lipídicas colesterol sérico total y triglicéridos y los marcadores renales ácido úrico y creatinina. El

índice de masa corporal (IMC) se ha definido como el cociente del peso evaluado en kilogramos dividido entre el cuadrado de la talla expresada en metros. Se ha considerado que un sujeto es obeso cuando su índice de masa corporal es igual o superior a 30 kg/m2. Se evaluaron también los valores de las tensiones arteriales sistólica y diastólica.

3.3. Análisis de potenciales factores de riesgo numéricos y categóricos.

En este epígrafe se analizan las asociaciones con la HTA de las variables numéricas edad, índice de masa corporal (IMC), colesterol sérico total, triglicéridos y ácido úrico y de los factores categóricos diabetes, obesidad (IMC categorizado), tabaquismo, sedentarismo e hipercreatinemia. En cada grupo de estudio, las variables numéricas se han resumido en medias y medias ajustadas por edad cuando se han verificado los supuestos de normalidad. En caso contrario, se han aplicado las transformaciones de Box y Cox para resumir finalmente los datos mediante las transformaciones inversas de las medias de los datos transformados. Las asociaciones de los factores binarios con la HTA se evalúan mediante las odd-ratios crudas. Utilizando los métodos de regresión logística, se ha determinado finalmente un conjunto de factores *independientes* asociados con la HTA. En la tabla 3.3.1 se exponen las variables observadas por población y sexo.

Se han resumido todas las afecciones cardiacas más comunes (véase 3.6.2) y cerebrales (véase 3.6.3). Para variables numéricas se muestran la media y la desviación típica global, además de las estimadas para hombres y mujeres.

Tabla 3.3.1. Recuento y descripción. Hipertensos y normotensos

	Hipertensos			
	Total n=368	Hombres n=162	Mujeres n=206	P
Edad	63.8 ± 13.0	64.2 ± 13.0	63.5 ± 12.9	0.531
Peso	79.5 ± 14.2	82.6 ± 14.0	77.9 ± 13.9	<0.001
Talla	1.64 ± 0.08	1.69 ± 0.07	1.59 ± 0.06	<0.001
IMC	29.7 ± 4.8	28.8 ± 4.1	30.4 ± 5.0	0.002
Obesidad	186 (50.5%)	70 (43.2%)	116 (56.3%)	0.013
Colesterol	238.3 ± 47.9	230.5 ± 44.8	244.5 ± 49.5	0.005
Triglicéridos	123 ×/ 1.52	123.4 ×/ 1.49	122.7 ×/ 1.55	0.906
Diabets	92 (25%)	37 (22.8%)	55 (26.7%)	0.396
Sedentarismo	215 (58.4%)	110 (53.4%)	105 (64.8%)	0.003
Tabaquismo	80 (21.7%)	63 (38.9%)	17 (8.3%)	< 0.001
Ácido úrico	5.10 ×/ 1.4	5.67×/1.0	4.70×/1.0	0.768
Hipercreatinemia	6 (1.6%)	4 (1.9%)	2 (1.2%)	0.595
Pat. cardiaca	102 (27.7%)	56 (34.5%)	46 (22.3%)	0.009
Pat. cerebral	32 (8.7%)	20 (12.3%)	12 (5.8%)	0.029
	Normotensos.			
	Total n=128	Hombres n=46	Mujeres n=82	P
Edad	57.1 ± 16.4	58.5 ± 17.4	56.3 ± 15.8	0.475
Peso	74 ± 12	78.8 ± 11.8	71.1 ± 11.3	0.008
Talla	1.62 ± 8.1	1.68 ± 5.8	1.59 ± 7.4	<0.001
IMC	28.9 ± 3.4	29 ± 3.51	28.81 ± 3.33	0.84.6
Obeso	46 (35.9%)	17 (37%)	29 (35.4)	0.748
Colesterol	232.6 ± 44.4	230.5 ± 51.3	233.9 ± 40.1	0.683
Triglicéridos	116 ×/ 158	125.6 ×/ 1.50	110.8 ×/ 1.62	0.148
Diabetes	10 (7.8%)	4 (8.7%)	6 (7.3%)	0.78
Sedentarismo	69 (53.9%)	14 (30.4%)	55 (67%)	0.023
Tabaquismo	101(78.9%)	43 (93.5%)	58 (70.7%)	0.002
Ácido úrico	4.91 ×/ 1.50	5.52×/1.3	4.60×/ 1.6	0.015
Pat. cardiaca	11 (8.6%)	8 (17.4%)	3 (3.6%)	0.008
Pat. cerebral	5 (3.9%)	3 (6.5%)	2 (2.4%)	0.253
Edad	57.1 ± 16.4	58.5 ± 17.4	56.3 ± 15.8	0.475
Peso	74 ± 12	78.8 ± 11.8	71.1 ± 11.3	0.008
Talla	1.62 ± 8.1	1.68 ± 5.8	1.59 ± 7.4	<0.001
IMC	28.9 ± 3.4	29 ± 3.51	28.81 ± 3.33	0.84.6
Obeso	46 (35.9%)	17 (37%)	29 (35.4)	0.748
Colesterol	232.6 ± 44.4	230.5 ± 51.3	233.9 ± 40.1	0.683
Triglicéridos	116 ×/ 158	125.6 ×/ 1.50	110.8 ×/ 1.62	0.148
Diabetes	10 (7.8%)	4 (8.7%)	6 (7.3%)	0.78
Sedentarismo	69 (53.9%)	14 (30.4%)	55 (67%)	0.023

El *p*-valor (P), hace referencia a la comparación de las medias de grupo mediante el *t*-test. Para los *triglicéridos* y el *ácido úrico* se ha tomado la escala logarítmica, y se han invertido los valores; por esta razón figura el doble signo ×/. Para las variables categóricas se exponen las prevalencias total y por grupo; en este caso, el *p*-valor (P) hace referencia al obtenido en el test de la ji-cuadrado. En la parte correspondiente a los normotensos, se ha omitido la variable hipercreatinemia por ser en este grupo de prevalencia cero.

Antes de empezar a mostrar los resultados de las diversas pruebas expuestas en el epígrafe 2.6.1 y 2.6.2, se realiza una primera exploración de datos, que incluye la representación P-plot y los contrastes de ajuste a la normalidad. Se aplica a cuatro variables numéricas (*colesterol, triglicéridos, glucemia y ácido úrico*). Aparece, para cada variable, un doble gráfico que corresponde a los datos crudos, y a los mismos tras la transformación de ByC.

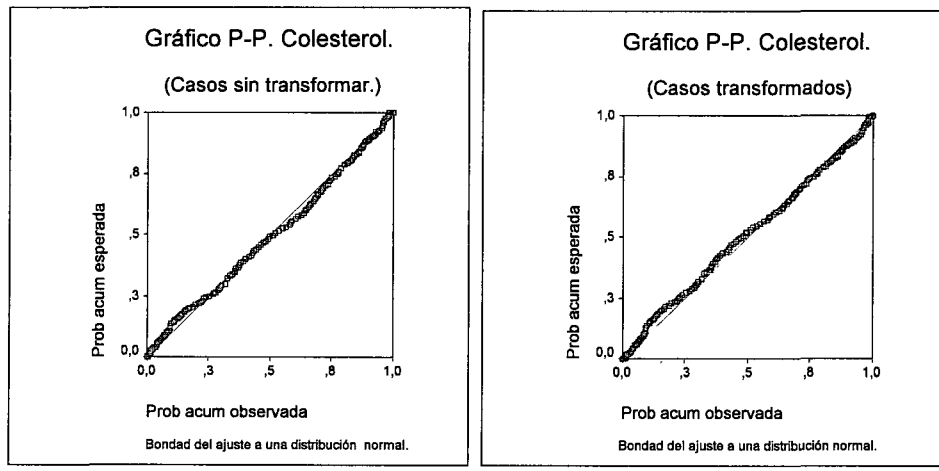


Fig. 3.3.2.

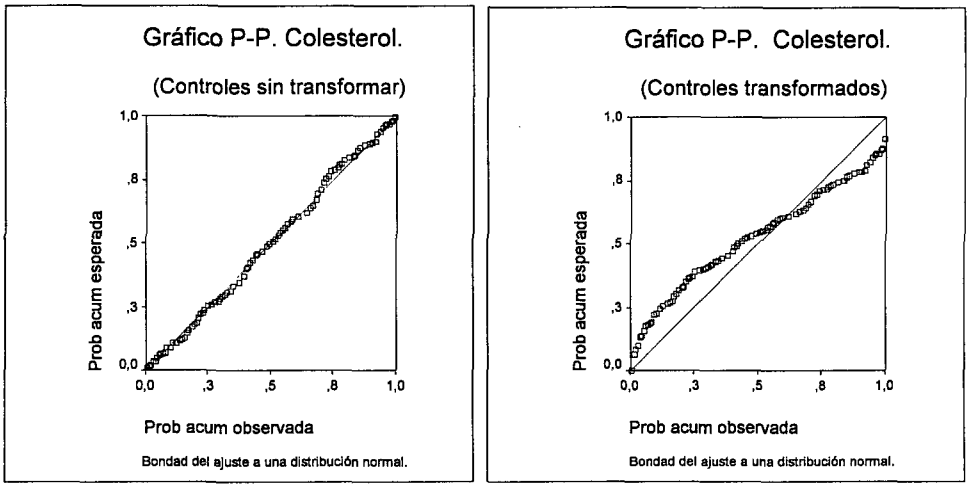


Fig. 3.3.2. (continua)

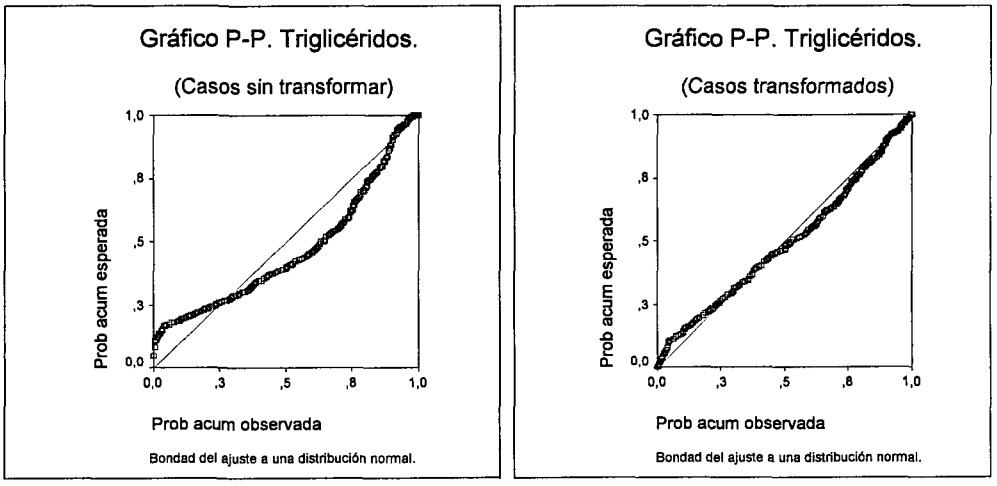


Fig. 3.3.3.

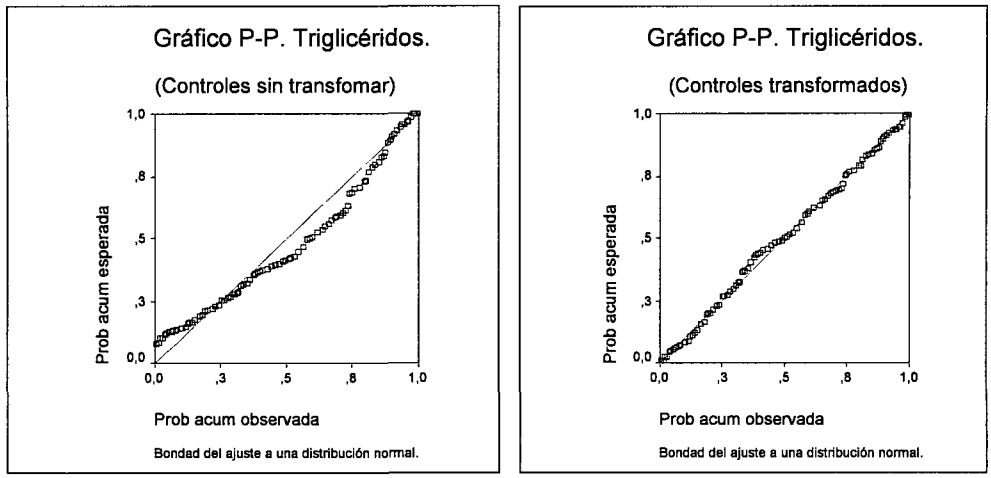


Fig. 3.3.3. (continua)

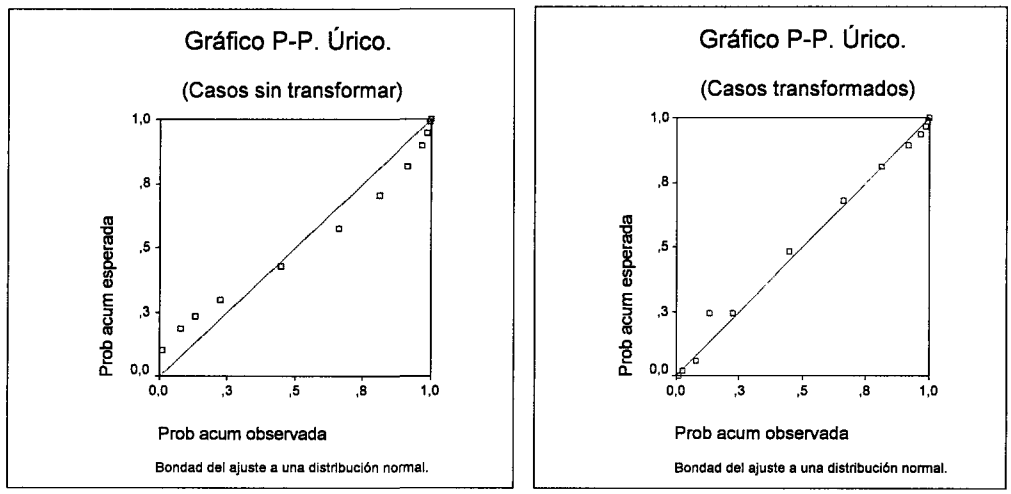
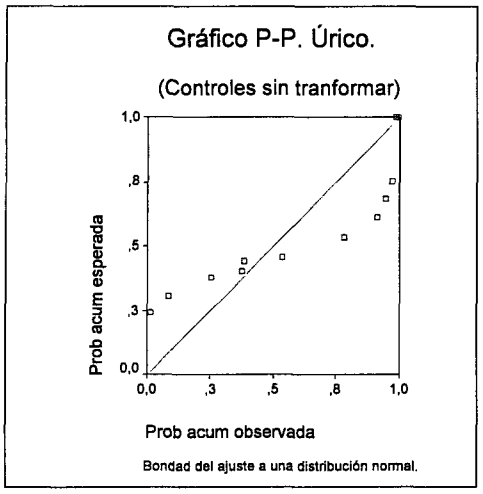


Fig. 3.3.4.



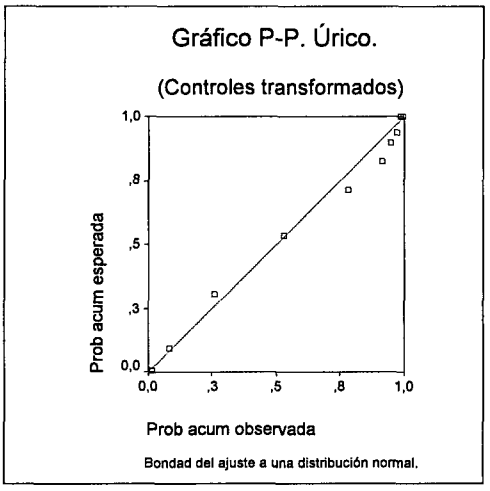
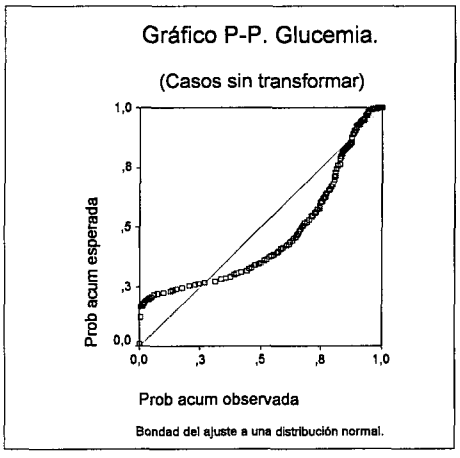


Fig 3.3.4. (continua)



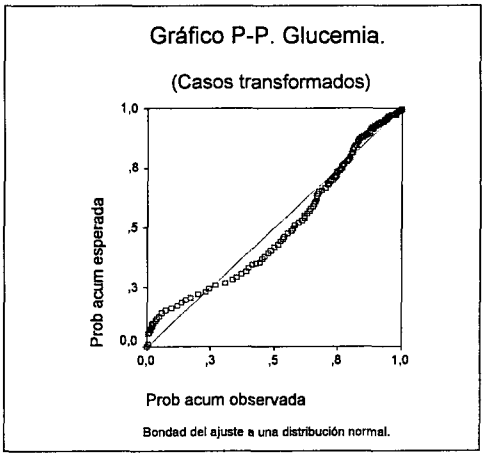


Fig 3.3.5.

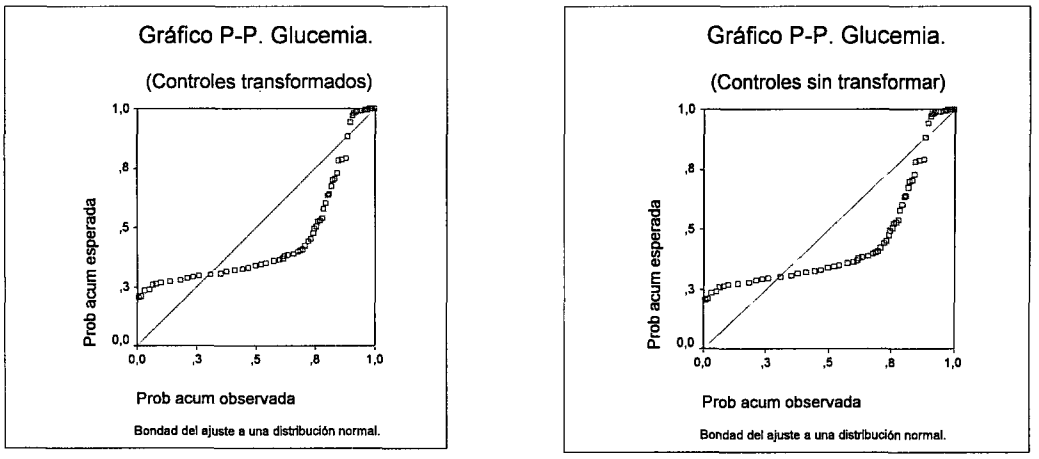


Fig. 3.3.5. (continua)

Como apreciación general sobre las figuras de 3.2.2 a 3.2.5 comentar que, basándonos solamente en este nivel gráfico, ya se hace evidente el *alejamiento* de la hipótesis de normalidad que presentan los datos originales, excepto en la variable *colesterol*. En general, mediante la transformación de ByC, se reduce el *alejamiento* de la distribución normal. En el caso de la variable *triglicéridos* se consigue normalidad.

La exploración de la normalidad de las variables numéricas analizadas y sus transformaciones de Box-Cox en cada uno de los grupos de estudios se resume en la tabla 3.3.6 en la cual se muestran los coeficientes de variación, medidas de forma (asimetría y curtosis), así como el test de normalidad de *Kolmogorov-Smirnov*. Se incluye también el test de Mann- Whitney para la comparación de las distribuciones de las variables entre ambos grupos de estudio. El test *K-S* contrasta la normalidad en cada grupo antes y después de la transformación.

Los resultados numéricos son indicativos, y están en perfecta consonancia con los gráficos. Así resulta evidente el espectacular acercamiento a normalidad de la variable *triglicéridos*, mientras que *úrico* apenas si disminuyen su distancia a

normalidad. Se recomienda, para una mejor comprensión de estas transformaciones, consultar las tablas con los datos originales y transformados para cada variable, y al mismo tiempo la parte gráfica correspondiente. Sobre las potencias λ, que figuran en tabla 3.3.7, cabe destacar el valor tan cercano a cero, aún más en el caso de la variable *triglicéridos* la potencia vale exactamente cero; lo que quiere decir que las transformaciones generalmente están cercanas a la logarítmica.

La variable *colesterol* presenta normalidad en los datos originales (Test *K-S,* $p=0.27$ y $p=0.90$ para casos y controles respectivamente). La transformada conjunta empeora los resultados de normalidad ($p=0.37$ y $p=0.01$), pues solo los casos conservan esta distribución. El contraste de posición de May-Withney ($p=0.35$ y $p=0.26$, original y transformado respectivamente), no detecta, en los datos originales ni transformados, diferencias significativas de posición; esto es, considera los datos extraídos de la misma población. Para esta variable, empeoran los índices de forma (asimetría y curtosis). Tras la transformación, los datos presentan una menor dispersión; este hecho se concreta en que las medias inversas y sus intervalos de confianza también son menores son menores que las originales. Un índice comparativo sobre la dispersión de la muestra es el Coeficiente de Variación cuyo valor en los datos originales es, respectivamente para casos y controles, 20% y 19%; pasando a valer en los transformados 0.7 % y 0.1% respectivamente, valores que expresan claramente la disminución de la dispersión. En este caso, la transformación mejora la dispersión, pero a costa de empeorar la forma y perder la normalidad. Por lo tanto, dependiendo de lo que se busque, se aplicará o no la transformación; si es normalidad el objetivo, no la aplicaremos, pues ya existe de por sí. Parecidas consideraciones podríamos hacer para las variables *edad* e *IMC.*

En la variable *triglicéridos* la transformación mejora la normalidad de una manera muy clara: en los datos originales el *contraste K-S* rechaza la hipótesis de normalidad ($p=0.01$, $p=0.05$, para casos y controles respectivamente), sin embargo en los datos transformados existe normalidad en todos los casos ($p=0.12$ y $p=0.91$).

El contraste de Mann-Withney (p=0.04, p=0.03, original y transformado respectivamente), detecta posicionamientos diferentes en ambas situaciones. Mejoran todas los índices de forma de un modo sensible. Las medias inversas son más pequeñas, menor su desviación típica y menores sus intervalos de confianza. Reduce considerablemente la dispersión, pues el Coeficiente de Variación es, en datos originales, para caso y controles de 58% y 52% respectivamente, pasando a ser el los transformados de 1.3% y de 1.5%. En este caso, la transformación no solamente consigue normalidad antes inexistente, sino que mejora claramente, en cuanto a forma y dispersión, los datos.

Para *ácido úrico* y *glucemia* no existe normalidad en los datos originales, ni tampoco se obtiene tras la transformación . El test de Mann-Withney detecta siempre un posicionamiento de los datos diferente. Mejoran las medidas de forma, asimetría y curtosis. Los estadísticos inversos indican una reducción importante de la dispersión; para hacer evidente este hecho, basta estimar los Coeficiente de Variación antes y después de la transformación. Las medias son más bajas y sus intervalos de confianza menores. No se consigue normalidad, pero la transformación mejora claramente la forma de la distribución y su dispersión.

Tabla 3.3.6. Forma y normalidad en datos originales y transformados.

		Asimetría.		Curtosis.		C.V.		P (K.S.)		M-Wth.
		HTA	NTA	HTA	NTA	HTA	NTA	HTA	NTA	
Edad	Original	-0.310	0.047	-0.025	-0.519	0.185	0.188	0.231	0.059	< 0.001
	Transformada	-1.522	-1.485	4.774	2.503	0.080	0.127	0.005	<0.001	< 0.001
IMC	Original	0.586	-0.334	0.644	-0.364	0.160	0.116	0.101	0.899	0.463
	Transformada	-0.006	-0.646	0.317	0.099	0.020	0.0159	0.070	0.693	0.458
Colesterol.	Original	0.372	0.22	1.28	-0.511	0.20	0.19	0.27	0.90	0.35
	Transformada	-0.54	-5.81	2.475	51.081	0.007	0.001	0.37	0.01	0.26
Triglicéridos	Original	2.34	1.58	9.68	3.81	0.58	0.52	0.01	0.05	0.04
	Transformada	0.70	-0.03	7.91	-0.09	0.013	0.015	0.12	0.91	0.03
Ácido úrico.	Original	9.31	6.81	137.27	49.26	0.50	0.90	0.01	0.01	0.03
	Transformada	-0.22	1.24	1.82	7.17	0.17	0.20	0.01	0.01	0.03
Glucemia.	Original	1.96	2.19	4.47	4.21	0.43	0.49	0.01	0.01	0.01
(λ =-0.8)	Transformada	-0.59	1.16	7.45	0.38	0.009	0.009	0.01	0.01	0.01

Tabla 3.3.7.Factores numéricos asociados a la HTA.

	ByC		Grupo		P	IC 95%
	λ		HTA n = 368	NTA n = 128		$\mu_D - \mu_C$
Edad	1	*o*	63.8 (62.5 – 65.1)	57.1 (54.2 – 59.9)	< 0.001 (*)	(3.92 – 9.53)
IMC	1	*o*	29.7 (29.2 – 30.3)	28.9 (27.9 – 29.8)	0.218 (*)	(-0.50 – 2.17)
		a	29.7 (29.2 – 30.2)	28.7 (27.5 – 29.9)	0.127 (***)	(-0.29 – 2.33)
Colesterol	1	*o*	238.3 (233.4 – 243.2)	232.6 (224.8 – 240.5)	0.241 (*)	(-3.84 – 15.24)
					0.240 (**)	(-3.48 – 14.91) (*b*)
		a	237.7 (232.9 – 242.5)	234.4 (226.2 – 242.8)	0.506 (***)	(-3.29 – 12.99)
Triglicéridos	0	*o*	155.7 (146.9 – 165.5)	136.0 (123.5 – 148.3)	0.020 (*)	(4.14 – 35.3) (*f*)
			(131.0) (μ)	(120.0) (μ)	0.026 (**)	(3.74 – 34.9) (*b*)
		t	137.0 (130.3 – 144.1)	120.4 (110.3 – 131.3)	0.011 (*)	(0.029 – 0.229)†
		a	136.3 (129.5 – 143.5)	122.2 (111.9 –133.35)	0.036 (***)	(0.007 – 0.211)†
Ácido úrico	-0.2	*o*	5.14 (4.95 – 5.33)	4.91 (4.56 – 5.28)	0.20 (*)	(-2.01 – 0.66) (*b*)
		t	5.08 (4.92 – 5.27)	4.83 (4.52 – 5.17)	0.20 (*)	(1.018 – 1.094)†
		a	5.12 (4.93 – 5.32)	4.96 (4.65 – 5.31)	0.428 (***)	(-0.008 – 0.094)†

(*o*) Medias observadas; (*t*) Imagen inversa de medias trasnformadas; (*a*) Medias ajustadas por edad; (μ) Mediana
(*) *t*-test; (**) Test de permutaciones bootstrap; (***) *F*-test.
$\lambda = 1$ indica que no se ha precisado transformación ByC. $\lambda = 0$ equivale a transformación logarítmica.
(*b*) Aproximación bootstrap a la distribución del pivotal
(*f*) Intervalo basado en la normalidad: puede ser poco admisible dada la evidente falta de normalidad de los datos.
(†) El intervalo de confianza es para la diferencia de medias en la escala transformada.

En la tabla 3.3.7 se muestran las variables numéricas que pudieran asociarse con HTA por grupo de estudio. Estas se han resumido en todos los casos en medias observadas o crudas (*o*), ajustadas por edad (*a*), y en aquellos casos que fue preciso, mediante la transformación inversa de la media de los datos transformados (*t*). En todos los casos, se utilizaron las transformaciones de Box y Cox (ByC) descritas en la sección 2.6. En aquellos casos en los que la exploración de datos permitió aceptar la hipótesis de normalidad, no se llevó a cabo tal transformación. En la segunda columna de la tabla se muestran los valores obtenidos para las correspondientes potencias de ByC. Cuando no se requirió realizar transformación, se consignó el valor $\lambda = 1$. En las columnas 3 y 4 se muestran las estimaciones de los parámetros de centralización correspondientes en los grupos de hipertensos y normotensos mediante intervalos de confianza al 95%. Los p-valores mostrados para la comparación de las medias crudas correspondientes a la edad e IMC se obtuvieron del t-test a dos colas. La normalidad de los datos previamente establecida garantiza el error *alpha* para estos contrastes. Para la edad solamente se estiman las medias crudas, sus intervalos de confianza al 95%, la significación del *t*-test y un intervalo de confianza para la diferencia de medias. Para el *IMC*, se añade el ajuste por la variable *edad*, dado que ésta pudiera actuar como factor de confusión. Es interesante notar que para el colesterol sérico total, el p-valor deducido del t-test ($p = 0.241$) y del test de permutaciones bootstrap ($p = 0.241$) son prácticamente idénticos. Ello es atribuible a la normalidad de esta variable en ambos grupos de estudio. Una valoración más detallada merece el análisis de los triglicéridos cuya exploración gráfica mediante el PP-plot (figura 3.3.3) muestra para esta variable un evidente distanciamiento de la normalidad, lo cual se confirma a través de los coeficientes de asimetría y curtosis (tabla 3.3.6), y sobre todo, por los contrastes de Kolmogorov-Smirnov para la normalidad dado en la misma tabla. Por este motivo, el *p*-valor obtenido para la comparación cruda de medias ($p = 0.020$) no garantiza en principio el error *alpha* de este contraste. No obstante, el efecto del teorema central del límite debido al alto tamaño muestral en cada uno de los grupos de estudio aporta una razonable fiabilidad a este p-valor. Nótese que éste es similar al dado por el test de permutaciones bootstrap (p = 0.026). El problema

mayor en esta comparación radica en el hecho de que las medias observadas (155.7 en el grupo HTA y 136.0 en el NTA) no resumen adecuadamente los dos conjuntos de datos. Se establece la hipertrigliceridemia, cuando esta variable supera el valor de 140. Obsérvese que este criterio entra en evidente conflicto con las medias referidas. Ello es atribuible precisamente a que cuando una distribución muestra largas colas a la derecha, la media no es un parámetro de centralización aceptable. La transformación de Box-Cox óptima es la logarítmica ($\lambda = 0$). Los parámetros de centralización así obtenidos (exponencial de la media de los logaritmos de los datos) son claramente aceptables, siendo éstos además similares a las medianas. Análoga valoración cabe hacer sobre el ácido úrico.

La diferencia entre las estimaciones de los parámetros de centralización de los triglicéridos según grupo de estudio podría ser atribuible a la hipertensión y/o a factores que concurren con ésta, tales como la diabetes mellitus. En la tabla 3.3.1 se muestran las prevalencia de esta enfermedad en ambos grupos de estudio (25% en HTA versus 7,8% en NTA). Es bien sabido que a menudo la diabetes concurre con elevación de los triglicéridos y depleción de la HDL (síndrome metabólico). Por tal motivo, la elevación de los triglicéridos podría en parte ser en parte atribuible a la diabetes, enfermedad ésta que se asocia con HTA. Por otra parte, la edad media en el grupo HTA (63.8 años) es significativamente superior a la del control (57.1 años), lo cual podría explicar también en parte la elevación de los triglicéridos. Por tal motivo, se muestran también las medias de esta variable ajustadas por edad mediante el siguiente modelo de análisis de la covarianza:

$$Y_{i,j} \cong N\left(\theta + \alpha_i + \beta \cdot \left(edad\right)_{i,j}, \sigma\right) : i = HTA, NTA\,;\, j = 1, \cdots, n_i$$

donde $Y_{i,j}$ representa la determinación de la variable (o una transformación de la misma), $\alpha_{HTA} = \alpha$ y $\alpha_{NTA} = 0$. Obviamente, la transformación de Box y Cox obtenida para la comparación entre los dos grupos de estudio que se han considerado en esta memoria no tiene por qué satisfacer las hipótesis de este modelo. No obstante, empíricamente podemos considerar la misma transformación (logarítmica para los triglicéridos) y una vez estimado el modelo,

comprobar si tales hipótesis se satisfacen mediante el análisis de residuales. La figura 3.3.8 muesta los residuales estudentizados representados frente a los valores pronósticos. Donde se observa la cercanía cero de muchos de ellos; salvo contados casos están en [-2,+2].

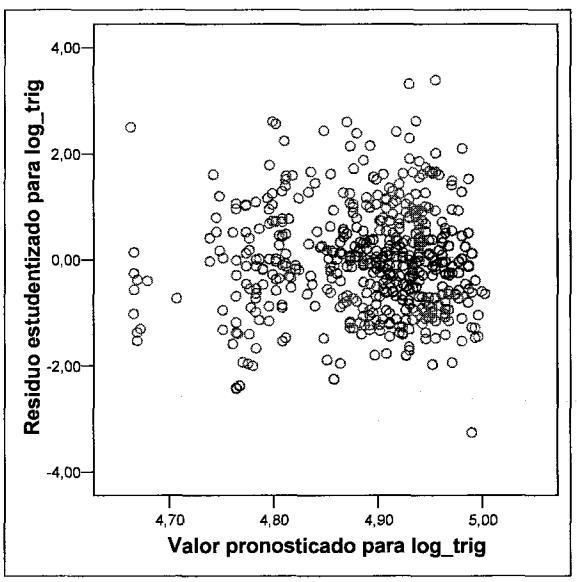


Fig. 3.3.8.

3.3.1. Resultados del análisis univariante de potenciales factores de riesgo categóricos.

Para resumir y simplificar algunos de los factores categóricos, y sin perder generalidad ni información que pudiera ser importante, se ha creado una nueva variable, hábitos sociales (*habsoc*), que resumen las variables *ejercicio, fuma e IMC*; pues, medidas sobre un mismo individuo, expresan modos de conducta sociosanitaria, con repercusiones sobre la salud. Esta nueva variable es binaria, con valor positivo, en referencia a una persona con un IMC no alto, que hace ejercicio y que no fuma. En caso contrario, el valor es negativo; esta modalidad es la contenida en la siguiente tabla .

De los resultados expresados en Tabla 3.3.9, que contiene una evaluación de posibles factores de riesgo categóricos para HTA, considerados en un análisis

preliminar que estima solo odds-ratios crudas, y ajustadas únicamente en caso de significación; se desprende la clara asociación existente entre *hta* y *diabetes* (p<0.001); su odd-ratio cruda vale 3.93, y su intervalo de confianza es (1.98 - 7.82). La ajustada es 2.78, siendo su IC (1.37 - 5.64).

Tabla 3.3.9..Evaluación de potenciales factores asociados a HTA.

	HTA		P	OR	
	Si (HTA) n=368	No (NTA) n=128		Cruda	Ajustada (Por *edad*)
Diabetes	25.0	7.8	<0.001	3.93 (1.98 - 7.82)	2.78 (1.37 – 5.64)
Obesidad	42.5	35.8	0.366	1.320 (0.72- 2.41)	
Tabaquismo	20.7	19.6	0.811	1.072(0.64 - 1.56)	
Sedentarismo	41.6	46.2	0.534	0.83 (0.46 - 1.49)	
Hábitos sociales	32.2	27.2	0.333	1.272(0.78 - 2.07)	
Hipercreatinemia	1.6	0	0.346		
Edad (por año)			<0.001		1.024 (1.01-1.04)

Es de señalar que ninguno de los IC contiene a la unidad, lo que indica que, no habiendo factor de confusión (la segunda odd-ratio ya está ajustada), es un verdadero factor de riesgo; observando la diferencia existente entre las dos odd-ratios, es evidente lo recomendable del ajuste. Consideraciones biológicas apuntan a que, como la diabetes supone un aumento de la glucemia, si realizásemos un análisis para esta última variable con HTA, obtendríamos un resultado significativo..

En el caso de la variable *creatinina*, no se estima odd-ratio por tener un valor de celda a cero. El resto de factores de riesgo no presentan p-valores significativos (p<0.05), y sus odd-ratios se encuentran próximas a la unidad.

Se incluye en Tabla 3.2.9. la variable *edad*, que resulta significativa (p<0.001), con una odd-ratio de 1.024 e IC (1.01 – 1.04); lo que supone que se trata de un verdadero factor de riesgo. No obstante, conviene recordar que se trata de un análisis preliminar y que todos estos potenciales factores de riesgo deben ser incluidos en un análisis multidimensional.

3.4. Estimación de curvas ROC.

Los resultados generados por la aplicación informática que implementa esta estimación (véase Apéndice II) se exponen a continuación. Se realizan tres estimaciones diferentes sobre los mismos datos, a saber: cruda (sin ningún tipo de suavizado); la segunda basada en una supuesta normalidad; y por último una estimación no paramétrica, usando estimadores de núcleo. En este último caso se ha usado el núcleo de Epanechnikov, para la obtención del bandwidth óptimo; se ha utilizado, además de la Regla del Pulgar, un método de validación cruzado como es el de la Máxima Verosimilitud; este método puede generar valores impropios de $ML(h)$, como es el caso, que por dos observaciones iguales, $\hat{f}_{h,i}(X_i)$ sea cero. Inconvenientes de este tipo, han sido salvados en la implementación.

En cada una de las metodologías empleadas proporciona la curva ROC y el área bajo la misma. Un valor cut-off para el criterio del máximo de la suma de sensibilidad y especificidad. Un segundo valor cut-off para el criterio de máxima sensibilidad. En este aspecto cabe recordar que este criterio es básicamente igual al anterior, únicamente se exige que la sensibilidad supere el 90%. Por último se construye un intervalo, basado en los valores Predictivo Positivo y Negativo.

La aplicación informa sobre el volumen de datos procesados, así como sobre valores relevantes usados en los cálculos.

3.4.1. Variable *Edad* como discriminante en el diagnóstico de HTA.

ESTIMACIÓN DE CURVAS ROC. R E S U L T A D O S.

Número de individuos procesados: 443
Número de individuos enfermos 350
Número de individuos sanos: 93
La prevalencia de la enfermedad, usada en los cálculos, es : 20.00 %

MÉTODO: ESTIMACIÓN CRUDA.

CRITERIO MAX.(ESP.+SENS): Cut-off = 65.00 ; Sensi.= 0.5272 ; Esp. = 0.7447
CRITERIO MAX. SENSI. : Cut-off = 46.00 ; Sensi. = 0.9284 ; Esp. = 0.1809
INTERVALO basado en Valor Pred.Pos.y Neg: Cut-off VPN 19.00 Prob. 0.9674
INTERVALO basado en Valor Pred.Pos y Neg: Cut-off VPP 65.00 Prob. 0.3405
ÁREA BAJO LA CURVA ROC : 0.62173

MÈTODO : ESTIMACIÓN NORMAL.
CRITERIO MAX.(ESP.+SENS): Cut-off 55.00 Sensi. 0.7642 Esp. 0.4286
CRITERIO MAX. SENSI. : Cut-off 47.00 ; Sensi. 0.9099 ; Esp. 0.2514
INTERVALO basado en Valor Pred.Pos.y Neg: Cut-off VPN 18.00 Prob. 0.9967
INTERVALO basado en Valor Pred.Pos y Neg: Cut-off VPP 69.00 Prob. 0.2589
ÁREA BAJO LA CURVA ROC : 0.60634

MÉTODO: ESTIMACIÓN NO PARAMETRICA. (Máxima verosimilitud)

CRITERIO MAX.(ESP.+SENS): Cut-off 66.00 ; Sensi. 0.5005 ; Esp. 0.7508
CRITERIO MAX. SENSI. : Cut-off 47.00 Sensi. 0.9172 Esp. 0.1797
INTERVALO basado en Valor Pred.Pos.y Neg: Cut-off VPN 18.00 Prob. 0.9749
INTERVALO basado en Valor Pred.Pos y Neg: Cut-off VPP 67.00 Prob. 0.3349 ÁREA BAJO LA CURVA ROC: 0.62139
Parámetros de Alisamiento: Enfermos = 1.117 y Sanos = 1.660

Resultados 3.4.1

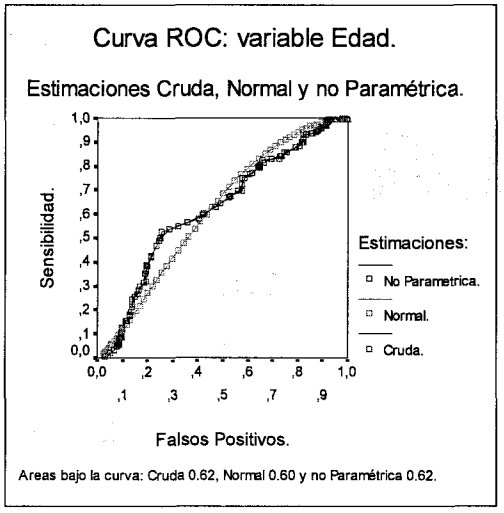


Fig 3.4.2.

La gráfica contenida en Fig. 3.4.2., representa superpuestas las curvas ROC de la triple estimación realizada para la variable *edad.*

3.4.2. Variable *Glucemia* como discriminante en el diagnóstico de HTA.

ESTIMACIÓN DE CURVAS ROC. R E S U L T A D O S.

Numero de individuos procesados: 494
Numero de individuos enfermos: 367
Numero de individuos sanos: 127
La prevalencia de la enfermedad, usada en los cálculos es: 20.00 %

MÉTODO: ESTIMACIÓN CRUDA.

CRITERIO MAX.(ESP.+SENS): Cut-off 101.00 Sensi. 0.6011 Esp. 0.6094
CRITERIO MAX. SENSI. : Cut-off 86.00 Sensi. 0.9235 Esp. 0.1641
INTERVALO basado en Valor Pred.Pos.y Neg: Cut-off VPN 64.00 Prob. 1.0000
INTERVALO basado en Valor Pred.Pos y Neg: Cut-off VPP 324.00 Prob. 0.3042
AREA BAJO LA CURVA ROC : 0.59914

MÉTODO: ESTIMACIÓN NORMAL.

CRITERIO MAX.(ESP.+SENS): Cut-off 96.00 Sensi. 0.7224 Esp. 0.3409
CRITERIO MAX. SENSI. : Cut-off 72.00 Sensi. 0.8485 Esp. 0.2061
INTERVALO basado en Valor Pred.Pos.y Neg: Cut-off VPN 72.00 Prob. 0.8448
INTERVALO basado en Valor Pred.Pos y Neg: Cut-off VPP 161.00 Prob. 0.2207
ÁREA BAJO LA CURVA ROC: 0.34675

MÉTODO : ESTIMACIÓN NO PARAMETRICA. (Máxima Verosimilitud)
CRITERIO MAX.(ESP.+SENS): Cut-off 107.00 Sensi. 0.5078 Esp. 0.6878
CRITERIO MAX. SENSI. : Cut-off 87.00 Sensi. 0.9018 Esp. 0.1750
INTERVALO basado en Valor Pred.Pos.y Neg: Cut-off VPN 72.00 Prob. 0.8302
INTERVALO basado en Valor Pred.Pos y Neg: Cut-off VPP 374.00 Prob. 1.00
ÁREA BAJO LA CURVA ROC : 0.6017

Parámetros de Alisamiento: Enfermos = 1.842 y Sanos = 2.661

Resultados 3.4.3.

La gráfica que viene a continuación, Figura.3.4.4, representa superpuestas las curvas ROC de las estimaciones cruda y no paramétrica realizada para la variable *glucemia*. Se omite la correspondiente a la estimación normal por su falta de interés.

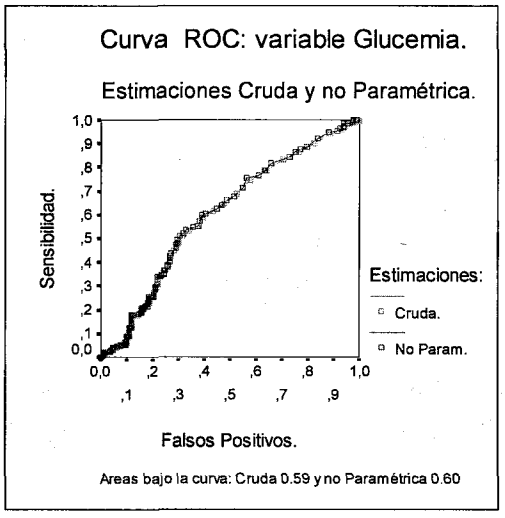


Fig. 3.4.4.

3.5. Marcadores multidimensionales: análisis discriminante.

A continuación figuran los resultados del análisis discriminante, cuyo desarrollo teórico se encuentra en el epígrafe 2.8. Se exponen un gráfico de dispersión de caso-control para la enfermedad en estudio, los coeficientes β_i $(i = 0,\cdots,3)$ de la función canónica discriminante; después se presenta la curva ROC correspondiente al marcador conjunto.

En el epígrafe 3.4.1 figura la función de clasificación lineal de Fisher y, por último, en epígrafe 3.4.2., el algoritmo de selección de variables para la discriminación. No se incluyen los resultados sobre valores predictivos, desarrollado en el epígrafe 2.8.3, pues desconocemos estimaciones de probabilidades a priori, necesarias para el desarrollo práctico de este punto.

La Fig. 3.4.1 proporciona una visión global de casos y controles, dando una primera idea de la situación de los mismos y posibilidades de discriminación.

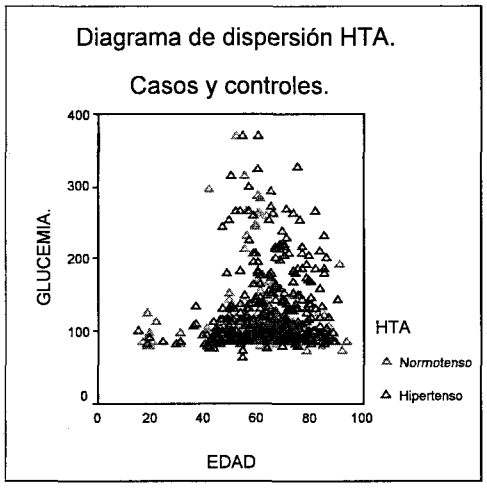


Fig. 3.5.1.

En la siguiente Tabla 3.4.2 figuran los coeficientes de la función canónica discriminante *Z*. Como sólo tenemos dos clases, normotenso e hipertenso, obtenemos una única a función discriminante.

Tabla 3.5.2.

Coeficiente de la función canónica discriminante Z.	
Cte.	$\beta_0 = -3.557$
Edad	$\beta_1 = 0.051$
Diabetes	$\beta_3 = 1.515$

La gráfica 3.5.3. representa la curva ROC para el marcador conjunto *Z*.

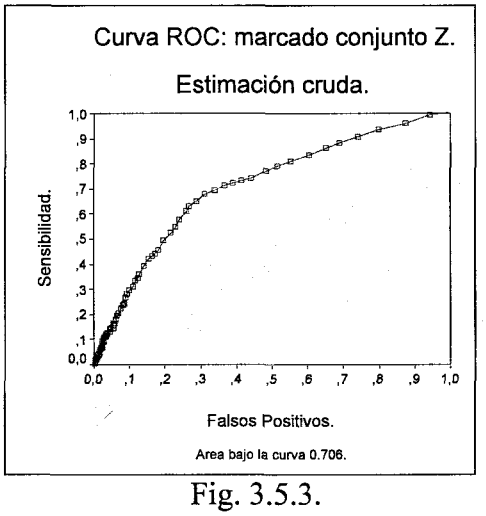


Fig. 3.5.3.

3.5.1 Funciones de clasificación lineal de Fisher.

La siguiente tabla contiene las funciones de clasificación de Fisher, una para la clase de los hipertensos y otra para normotensos.

Tabla 3.5.4.

Coeficientes de la función de clasificación

	tipo	
	Normotenso	Hipertenso
EDAD	,315	,342
diabetes	-,705	,077
(Constante)	-9,885	-11,644

Funciones discriminantes lineales de Fisher

En Tabla 3.4.5. figura la clasificación promovida por las funciones de Fisher.
La clasificación dada por estas funciones discriminantes de Fisher está contenida en Tabla 3.4.5., donde en filas tenemos, para normotenso e hipertenso respectivamente, los originales, en columnas los pronosticados. Las coincidencias en fila y columna son casos bien clasificados.

Tabla 3.5.5.

Resultados de la clasificación[a]

			Grupo de pertenencia pronosticado		
		tipo	Normotenso	Hipertenso	Total
Original	Recuento	Normotenso	94	34	128
		Hipertenso	161	207	368
	%	Normotenso	73,4	26,6	100,0
		Hipertenso	43,8	56,3	100,0

a. Clasificados correctamente el 60,7% de los casos agrupados originales.

3.5.2 Selección de variables para la discriminación.

Para el análisis se parte de la siguiente pila de potenciales variables discriminantes*: edad, sexo, habsoc, coles, glucem, trigl, diabetes, úrico.* Con un total de individuos *n=496.* Se aplica el algoritmo contenido en el epígrafe 2.8.4.

En Tabla 3.4.6 se encuentra resumida la ejecución del algoritmo para la selección de variables. Solamente se procesan 443 individuos, pues el resto hasta 496 tiene perdido algún valor de las variables discriminantes, por esta razón los grados de libertad (*gl2*) del *F*-test comienzan en 441.

Tabla 3.5.6. Selección de variables.

Variables introducidas/eliminadas[a,b,c,d]

Paso	Introducidas	Mín. F				
		Estadístico	gl1	gl2	Sig.	Entre grupos
1	EDAD	11,879	1	441,000	,001	Normotenso y Hipertenso
2	diabetes	9,764	2	440,000	7,092E-05	Normotenso y Hipertenso

En cada paso se introduce la variable que maximiza la razón F menor entre pares de grupos.

a. El número máximo de pasos es 16.

b. La F parcial mínima para entrar es 3.84.

c. La F parcial máxima para eliminar es 2.71

d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

3.6. Resultados regresión logística: determinación de factores de riesgo.

La metodología de regresión logística ya se ha expuesto anteriormente, pero debemos significar la novedad histórica de la misma, desarrollada casi enteramente en la década de los ochenta, y posteriormente implementada en diversos paquetes estadísticos.

Los paquetes BMDP, SAS, SPSS entre otros, tienen implementada esta metodología. En este caso todo el proceso computacional ha sido desarrollado con el programa SPSS 11.0, empleando el procedimiento de regresión logística binaria. De entre los métodos concretos empleados, cabe destacar que el algoritmo de selección empleado es el de la Razón de Verosimilitudes, es su modalidad por pasos hacia delante o forward (parte de un modelo con solo una constante, y según puntuación de variables, estadístico y p-valor asociado, las variables entran en el modelo siempre y cuando $p<0.05$).

Por otro lado, en este imprescindible contexto común, debe figurar la variable *edad*, de singular importancia en ciencias de la vida, estimando las odd-ratios ajustados por la misma. Estos resultados ya se encuentran incluidos en la tabla 3.3.9, aunque en este epígrafe se muestra el proceso desarrollado.

Todas las variables excepto, *edad*, son categóricas con lo que se expresan en forma Dummy (0=valor1,1=valor2), cuestión necesaria tanto por razones de computación como por razones biológicas ya explicadas.

Los resultados expresados en el epígrafe 3.6.1, proceden de la aplicación de la metodología implementada en SPSS a la variable *hta*, considerándola como enfermedad; este es uno de los principales objetivos de esta memoria. Las aplicaciones a las afecciones cardiacas y de cerebro, recogidas en los epígrafes 3.6.2 y 3.6.3 respectivamente, reflejan la aplicación del método a estas enfermedades.

En cada caso se han seguido los pasos básicos, aunque no siempre se exponen con el mismo detalle. En los siguientes apartados hay algunos contenidos diferentes; esto se hace por considerar algunas cuestiones redundantes y para poder dar resultados de todos los apartados que componen la metodología. En el epígrafe 3.6.2 figura la tabla de clasificación de la enfermedad. El criterio de información de Akaike (ACI) no se encuentra implementado en el paquete, pero se incluye por considerarlo de interés.

Se intenta dar una panorámica de resultados exhaustiva, pero sin caer en redundancia, de ahí que en cada ejemplificación se exponga, además de las cuestiones básicas, lo más destacable de la misma.

Antes de desarrollar en detalle la regresión logística para las tres variables *hta*, *corazón* y *cerebro*, resumimos los resultados en la Tabla 3.6.1. De los tres modelos, solamente el estimado para hipertensión no tiene valor predictivo, en los otros dos *hta* es solamente una variable que expresa un potencial factor de riesgo.

Tabla 3.6.1. Resumen modelos de Regresión Logística. OR crudas y ajustadas.

	Variables, odd-ratio crudo y ajustado e IC al 95%					
	Cte.	*Edad.*	*Sexo.*	*Hta.*	*Diabetes.*	**HyL(P)**
Corazón.		OR_C=1.040	OR_C=2.168	OR_C=4.079	OR_C=2.837	
		(1.02-1.06)	(1.42-3.32)	(2.11-7.88)	(1.77-4.55)	
		OR_a=1.035	OR_a=2.255	OR_a=2.567	OR_a=1.151	
	-4.750	(1.01-1.05)	(1.42-3.57)	(1.25-5.27)	(1.30-3.56)	0.69
Cerebro.		OR_C=1.055	OR_C=2.433		OR_C=3.305	
		(1.02-1.09)	(1.22-4.85)		(1.65-6.60)	
		OR_a=1.078	OR_a=4.996		OR_a=5.334	
	-9.390	(1.03-1.12)	(1.98-2.61)		(2.30-2.37)	0.237

3.6.1. Modelo de regresión logística para HTA.

Regresión logística: variable dependiente *hta*.

Cantidad de información de Akaike.

Inicialmente (r=2,p=8).

ACI=461.496 + 20 =481.496

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	1.304	.116	127.063	1	.000	3.684

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	**EDAD**	**11.610**	1	**.001**
		SEXO	2.065	1	.151
		HABSOC	.518	1	.472
		URICOBIN	2.718	1	.099
		TRIBIN	1.139	1	.286
		COLESBIN	.476	1	.490
		DIABETES	10.504	1	.001
		CREAT	1.646	1	.199
	Estadísticos globales		24.811	8	.002

Entra en el modelo la variable edad, pues su puntuación es la más alta y su *p* - valor el más bajo ($p<0.05$).

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 1	Variables	SEXO	1.602	1	.206
		HABSOC	1.178	1	.278
		URICOBIN	2.267	1	.132
		TRIBIN	.840	1	.359
		COLESBIN	1.014	1	.314
		DIABETES	**8.010**	1	**.005**
		CREAT	1.411	1	.235
	Estadísticos globales		13.797	7	.055
Paso 2	Variables	SEXO	1.880	1	.170
		HABSOC	.636	1	.425
		URICOBIN	2.124	1	.145
		TRIBIN	.334	1	.563
		COLESBIN	.742	1	.389
		CREAT	1.248	1	.264
	Estadísticos globales		6.022	6	.421

Introducida la variable edad en el modelo, vemos como el resto de variables en general disminuyen su puntuación y aumenta su significación. Esto se aprecia de un modo muy claro en la variable uricobin lo que significa que la edad era un factor de confusión.

Entra el modelo la variable diabetes (su *p*-valor es el más bajo e inferior a 0.05).

Cantidad de información de Akaike.

Vuelve a disminuir.

Paso 1: ACI=450.216 + 20 = 470.216

Paso 2: ACI=441.259 + 20 = 461.259

No entran nuevas variables en el modelo, pues ningún *p*-valor es significativo.

Tabla 3.6.2.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95.0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	EDAD	.028	.008	11.163	1	.001	1.028	1.011	1.045
	Constante	-.385	.508	.574	1	.449	.680		
Paso 2[b]	**EDAD**	.024	.008	8.180	1	.004	**1.024**	**1.007**	**1.041**
	DIABETES	.987	.360	7.527	1	.006	**2.683**	**1.326**	**5.430**
	Constante	-.319	.506	.396	1	.529	.727		

a. Variable(s) introducida(s) en el paso 1: EDAD.

b. Variable(s) introducida(s) en el paso 2: DIABETES.

Historial de iteraciones[a,b,c,d,e]

Iteración		-2 log de la verosimilitud	Coeficientes		
			Constante	EDAD	DIABETES
Paso 1	1	453.672	-.055	.019	
	2	450.237	-.350	.027	
	3	**450.216**	-.385	.028	
Paso 2	1	446.939	-.027	.017	.517
	2	441.406	-.286	.023	.874
	3	441.260	-.318	.024	.981
	4	**441.259**	-.319	.024	.987

a. Método: Por pasos hacia adelante (Razón de verosimilitud)

b. En el modelo se incluye una constante.

c. -2 log de la verosimilitud inicial: 461.496

d. La estimación ha finalizado en el número de iteración 3 porque el logaritmo de la verosimilitud ha disminuido en menos de un .010 por ciento.

e. La estimación ha finalizado en el número de iteración 4 porque el logaritmo de la verosimilitud ha disminuido en menos de un .010 por ciento.

Eliminando la edad en el paso 1, obtenemos que la significación del cambio (estadístico *G*) es $p<0.01$, por lo que la misma resulta significativa, en cuanto a factor de riesgo de Hta.

Lo mismo ocurre en el paso 2 con edad y diabetes ($p<0.01$ en ambos casos).

A continuación se expone el test de Hosmer y Lemeshow.

Tabla 3.6.3.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	13.136	8	.107
2	15.408	8	.052

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		tipo = Normotenso		tipo = Hipertenso		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	17	15.338	26	27.662	43
	2	9	11.987	34	31.013	43
	3	14	10.467	28	31.533	42
	4	8	10.409	38	35.591	46
	5	16	10.125	32	37.875	48
	6	8	7.134	28	28.866	36
	7	4	7.799	38	34.201	42
	8	5	7.870	41	38.130	46
	9	5	7.072	41	38.928	46
	10	9	6.806	44	46.194	53
Paso 2	1	16	17.280	31	29.720	47
	2	14	13.106	30	30.894	44
	3	7	10.661	33	29.339	40
	4	18	12.155	32	37.845	50
	5	10	9.519	32	32.481	42
	6	5	8.946	38	34.054	43
	7	6	8.194	38	35.806	44
	8	12	6.768	33	38.232	45
	9	6	4.999	41	42.001	47
	10	1	3.371	42	39.629	43

Tabla 3.6.4.

Listado de residuos por casos

Seleccionados.		Observados	Pronosticados	G P	Residuo.	
Caso		Tipo			Resd.	ZResid.
379	S	N**	.896	H	-.896	-2.938
381	S	N**	.880	H	-.880	-2.705
382	S	N**	.887	H	-.887	-2.802
382	S	N**	.887	H	-.887	-2.802
385	S	N**	.869	H	-.869	-2.580
387	S	N**	.880	H	-.880	-2.705
390	S	N**	.887	H	-.887	-2.802
391	S	N**	.892	H	-.892	-2.869
394	S	N**	.916	H	-.916	-3.306
437	S	N**	.864	H	-.864	-2.526
438	S	N**	.867	H	-.867	-2.556
445	S	N**	.869	H	-.869	-2.580

a S = Seleccionados, N = Casos no seleccionados y ** = Casos mal clasificados. GP=Grupo Pronosticado. b Se listan los casos con residuos tipificados mayores que 2.000 la SD (Criterio de HyL)

Si nos fijamos en Tabla 3.6.4., el residual más grande es el correspondiente al individuo 394 con un valor de –3.306. Eliminado este individuo y realizado un nuevo procesamiento se obtiene:

Las variable en la ecuación son las mismas. La edad en idénticos términos y la diabetes tiene de odd-ratio e intervalo de confianza respectivamente, 2.979 e IC (95%) [1.429, 6.210]. Este nuevo modelo tiene un *ACI* final de 456.200 y, realizado es test de HyL, con 8 grados de libertad, la significación es $p<0.017$. Los nuevos residuos originados por este modelo, se mantienen, en términos absolutos, menores que 3.

Resultado 3.6.5.

3.6.2. Modelo de regresión logística para afecciones cardiacas.

Resulta de interés conocer los factores que intervienen en grupos de enfermedades presuntamente relacionadas con HTA, y que afectan a un determinado grupo de patologías, y que se centran sobre un determinado órgano o función, como pueden ser las de corazón. En la base de datos están recogidas: angina, infarto, hipertrofia ventricular izquierda e insuficiencia cardiaca, que si bien no son exhaustivas, si se encuentran entre la más frecuentes.

Consideramos que una persona ha sufrido alguna afección de corazón si ha padecido alguna de las enfermedades mencionadas anteriormente. A esta variable le llamaremos ***corazón*** (codificado: 0=negativo, 1=afirmativo), estructurando para ella un modelo de regresión logística que analice los factores de riesgo que intervienen en ella, sus respectivas odds ratios e intervalos de confianza. En este caso como la variable *hta*, figura como una de las posibles variables predictoras, el modelo obtenido tiene carácter predictivo. Partimos de las posibles variables predictoras: *edad, sexo, habsoc, colesbin, tribin, glucebin, uricobin, diabetes y hta*. Damos la salida, comentada, elaborada por el programa SPSS 11.0.

Regresión logística: variable dependiente corazón

Tabla 3.6.6

Variables en la ecuación

		B	E.T.	Wald		Sig.	Exp (B)	I.C. 95,0% para EXP(B)	
								Inf.	Sup.
Paso 1[a]	EDAD	,038	,009	17,18	1	,000	1,038	1,020	1,057
	Constante	-3,534	,612	33,38	1	,000	,029		
Paso 2[b]	EDAD	,037	,009	16,12	1	,000	1,038	1,019	1,056
	SEXO	,761	,227	11,24	1	,001	2,141	1,372	3,341
	Constante	-3,845	,630	37,19	1	,000	,021		
Paso 3[c]	EDAD	,034	,009	13,18	1	,000	1,035	1,016	1,054
	SEXO	,808	,231	12,20	1	,000	2,244	1,426	3,531
	DIABETES	,849	,253	11,30	1	,001	2,337	1,425	3,834
	Constante	-3,931	,649	36,69	1	,000	,020		
Paso 4[d]	EDAD	,032	,010	11,09	1	,001	1,033	1,013	1,053
	SEXO	,785	,233	11,31	1	,001	2,192	1,388	3,464
	DIABETES	,748	,256	8,512	1	,004	2,113	1,278	3,492
	HTA	,965	,366	6,944	1	,008	2,625	1,281	5,380
	Constante	-4,568	,719	40,32	1	,000	,010		

a. Variable(s) introducida(s) en el paso 1: EDAD.

b. Variable(s) introducida(s) en el paso 2: SEXO.

c. Variable(s) introducida(s) en el paso 3: DIABETES.

d. Variable(s) introducida(s) en el paso 4: HTA.

Tabla 3.6.7.

Cantidad de información de Akaike.

Inicialmente (r=2,p=8).

ACI=499.928 + 20 = 519.928

Las cantidades de información van disminuyendo a cada paso, lo que indica que cada nueva variable en el modelo aporta información.

Paso2

ACI=480.757 + 20 = 500.757

Paso3

ACI=469.379 + 20 = 489.379

Paso4

ACI=458.322 + 20 = 478.322

Paso 5 y último.

ACI=450.221 + 20 = 470.221

Tabla 3.6.8.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	12,099	8	,147
2	17,289	8	,027
3	9,593	8	,295
4	5,490	8	,704

Tabla 3.6.9. **Listado de residuos por casos**

	Seleccionados.	Observados	Pronosticados	G P	Residuo.	
Caso		Tipo			Resd.	ZResid.
192	S	S**	,117	N	,883	2,752
227	S	S**	,098	N	,902	3,031
283	S	S**	,082	N	,918	3,339
312	S	S**	,098	N	,902	3,031
369	S	S**	,097	N	,903	3,060
382	S	S**	,128	N	,872	2,611
388	S	S**	,067	N	,933	3,734
398	S	S**	,124	N	,876	2,663
480	S	S**	,039	N	,961	4,961

a S = Seleccionados, N = Casos no seleccionados y ** = Casos mal clasificados.
b Se listan los casos con residuos tipificados mayores que 2.000.

Visto que se originan residuos grandes, de valores superiores a 3 e incluso uno de valor superior a 4, resultaría recomendable eliminar los mayores y realizar una nueva RL prescindiendo de estos casos. Los resultados de la misma, eliminando los casos 480 y 283 se exponen a continuación.

Regresión logística: variable dependiente corazón. Modelo depurado.

Tabla 3.6.10

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	EDAD	,040	,009	18,662	1	,000	1,041	1,022	1,060
	Constante	-3,693	,624	35,085	1	,000	,025		
Paso 2[b]	EDAD	,039	,009	17,629	1	,000	1,040	1,021	1,059
	SEXO	,787	,229	11,846	1	,001	2,196	1,403	3,437
	Constante	-4,025	,644	39,090	1	,000	,018		
Paso 3[c]	EDAD	,037	,010	14,630	1	,000	1,037	1,018	1,057
	SEXO	,836	,233	12,863	1	,000	2,307	1,461	3,643
	DIABETES	,866	,254	11,644	1	,001	2,377	1,446	3,909
	Constante	-4,127	,664	38,655	1	,000	,016		
Paso 4[d]	EDAD	,035	,010	12,499	1	,000	1,035	1,016	1,056
	SEXO	,813	,235	11,966	1	,001	2,255	1,423	3,575
	DIABETES	,766	,258	8,847	1	,003	2,151	1,299	3,564
	HTA	,943	,367	6,587	1	,010	2,567	1,250	5,272
	Constante	-4,750	,734	41,922	1	,000	,009		

a. Variable(s) introducida(s) en el paso 1: EDAD.

b. Variable(s) introducida(s) en el paso 2: SEXO.

c. Variable(s) introducida(s) en el paso 3: DIABETES.

d. Variable(s) introducida(s) en el paso 4: HTA.

Tabla 3.6.11.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	12,753	8	,121
2	17,956	8	,022
3	8,112	8	,423
4	5,607	8	,691

Tabla 3.6.12.

Cantidad de información de Akaike.

Las cantidades de información van disminuyendo, lo que indica que cada nueva variable en el modelo aporta información.

Paso1

ACI=476.183 + 20 = 496.183

Paso 5 y último.

ACI=445.117 + 20 = 465.117

Tabla 3.6.13.

Tabla de clasificación[a]

Observado			Pronosticado		
			CORAZON		Porcentaje correcto
			No	Si	
Paso 1	CORAZON	No	333	1	99,7
		Si	110	0	,0
	Porcentaje global				75,0
Paso 2	CORAZON	No	328	6	98,2
		Si	103	7	6,4
	Porcentaje global				75,5
Paso 3	CORAZON	No	321	13	96,1
		Si	99	11	10,0
	Porcentaje global				74,8
Paso 4	CORAZON	No	319	15	95,5
		Si	98	12	10,9
	Porcentaje global				74,5

a. El valor de corte es ,500

Tabla 3.6.14. **Listado de residuos por casos.**

	Seleccionados.	Observados	Pronosticados	G P	Residuo.	
Caso		Tipo			Resd.	ZResid.
191	S	S**	,135	N	,865	2,535
192	S	S**	,109	N	,891	2,863
216	S	S**	,131	N	,869	2,580
225	S	S**	,131	N	,869	2,580
227	S	S**	,090	N	,910	3,178
311	S	S**	,090	N	,910	3,178
368	S	S**	,094	N	,906	3,108
381	S	S**	,126	N	,874	2,628
387	S	S**	,065	N	,935	3,789
397	S	S**	,126	N	,874	2,630

a S = Seleccionados, N = Casos no seleccionados y ** = Casos mal clasificados.
b Se listan los casos con residuos tipificados mayores que 2.000.

La siguiente gráfica de dispersión muestra , ΔX^2, en función de las probabilidades pronosticadas por el modelo, $\hat{\pi}$. Esto es, los pares ($\hat{\pi}, \Delta X^2$) (véase epígrafe 2.9.5.4.).(En Fig. 3.6.15. (a), el eje vertical tiene la leyenda DELTA, en referencia a DELTA (X) CUADRADO)

GRAFICA: ΔX^2, versus $\hat{\pi}$. (Modelo sin depurar)

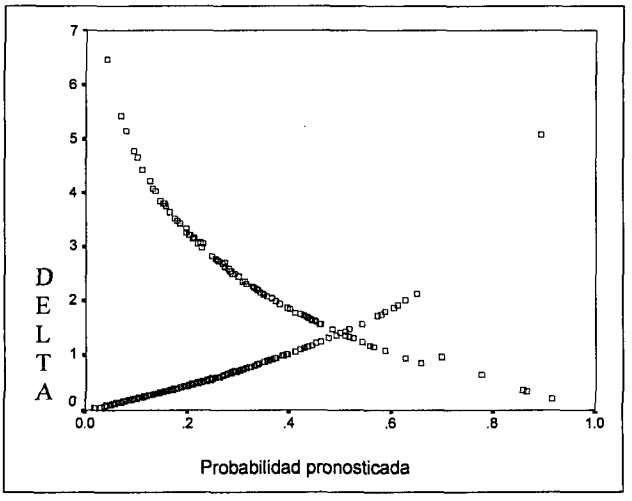


Fig. 3.6.15. (a)

GRAFICA: ΔX^2, versus $\hat{\pi}$. (Modelo depurado)

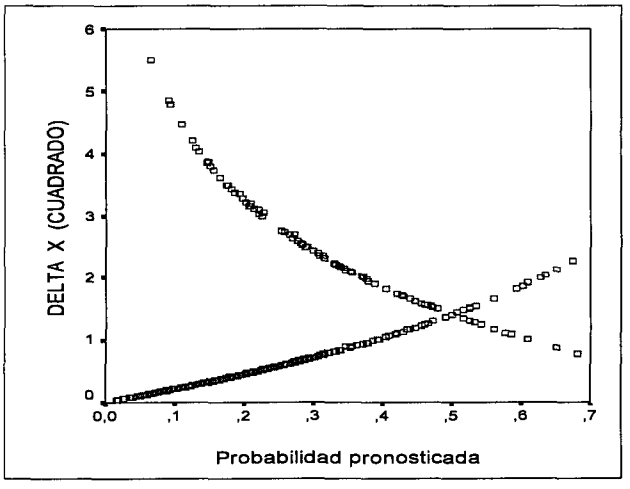


Fig. 3.6.15. (b)

El arco de curva que tiene su mínimo cercano a cero y es monótono creciente, representa el patrón de respuesta Y=0; mientras que el otro arco, monótono decreciente, con mínimo cercano a 0.7, representa el patrón de respuesta Y=1. La ordenada de los puntos es proporcional a $(1-\hat{\pi}_j)^2$ o bien $(0-\hat{\pi}_j)^2$ $\forall x_j$, para patrones Y=1 e Y=0, respectivamente.

3.6.3. Modelo de regresión logística para afecciones de cerebro.

Analizamos ahora un grupo de enfermedades que afectan al cerebro, presuntamente relacionadas con HTA. En la base de datos están recogidas accidente cerebro vascular, hemorragia cerebral y trombosis, que si bien no son exhaustivas, si se encuentran entre la más frecuentes. Consideramos que una persona ha sufrido alguna afección de cerebro, si ha padecido alguna de las enfermedades mencionadas anteriormente. A esta variable le llamaremos ***cerebro*** (codificado: 0=negativo, 1=positivo), estructurando para ella, un modelo de regresión logística que analice los factores de riesgo que intervienen en la misma, sus respectivas odd-ratios e intervalos de confianza. En este caso como la variable *hta*, figura como una de las posibles variables predictoras, el modelo obtenido tiene carácter predictivo.

Partimos de las posibles variables predictoras: *edad, sexo, habsoc, colesbin, tribin, glucebin, diabetes y hta.*

Damos la salida, resumida, elaborada por el programa SPSS 11.0.

Regresión logística: variable dependiente cerebro.

Tabla 3.6.16.

Cantidad de información de Akaike.
Las cantidades de información van disminuyendo a cada paso, lo que indica que cada nueva variable en el modelo, aporta información. **Inicialmente (r=2, p=8)** **ACI=208.697+20= 228.697** **Paso3 y ultimo** **ACI=162.896 + 20 =182.896**

Tabla 3.6.17

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	DIABETES	1,626	,401	16,478	1	,000	5,085	2,319	11,151
	Constante	-3,308	,294	126,689	1	,000	,037		
Paso 2[b]	SEXO	1,623	,460	12,449	1	,000	5,068	2,057	12,485
	DIABETES	1,746	,413	17,829	1	,000	5,729	2,548	12,883
	Constante	-4,299	,464	86,016	1	,000	,014		
Paso 3[c]	EDAD	,075	,021	12,700	1	,000	1,078	1,034	1,123
	SEXO	1,609	,472	11,596	1	,001	4,996	1,979	12,612
	DIABETES	1,674	,429	15,224	1	,000	5,334	2,301	12,368
	Constante	-9,390	1,625	33,373	1	,000	,000		

a. Variable(s) introducida(s) en el paso 1: DIABETES.

b. Variable(s) introducida(s) en el paso 2: SEXO.

c. Variable(s) introducida(s) en el paso 3: EDAD.

Tabla 3.6.18.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
2	5,861	2	,053
3	10,424	8	,237

El modelo probabilístico, en términos matemáticos, es:

$$P(cerebro \mid edad, sexo, diabetes) = \frac{\exp(-9.390 + 0.75(edad) + 1.609(sexo) + 1.674(diabetes)}{1 + \exp(-9.390 + 0.75(edad) + 1.609(sexo) + 1.674(diabetes)}$$

Resultado 3.6.19.

Donde la variable *edad* $\in$ N. Y las variables *sexo, diabetes* están codificada 0,1.

3.6.4. Corrección de la sobreestimación del vector beta de coeficientes de regresión logística.

Los resultados obtenidos mediante metodología bootstrap, aplicada a la base de datos descrita en el epígrafe 3.1, tanto para la estimación de *Factores de Corrección Lineal*, como para observar el número de variables contenidas en el modelo, se exponen en los siguientes epígrafes 3.6.4.1. y 3.6.4.2. respectivamente.

Parece necesario comentar que el programa informático, que implementa estas aplicaciones, se encuentra en el Apéndice II de esta memoria; se trata de una herramienta creada para resolver este problema. Los resultados han sido doblemente contrastados con el paquete SPSS 11.0 y librerías Fortran.

3.6.4.1. Sobredimensionamiento del vector beta.

En Tabla 3.6.20. se muestra la variación de los coeficientes, según el número de casos n. Se designa el número de variables por p, consideramos p=9 (*edad, sexo, hábitos sociales, colesterol, triglicéridos, ácido úrico, glucemia, diabetes* y *creatinina*).

Según las cantidades $n>>p$ o bien, violando claramente la regla uno diez, esto es, $n<10p$; se obtienen los siguientes resultados:

Tabla 3.6.20.

N° Casos	β_0 (Cte.)	SE	β_1 (Ed.)	SE	β_2 (Sex)	SE	β_8 (Diabt.)	SE	Vars
n=443	-.3122	.5103	.0239	.0083			.9608	.3602	p=2
n=86	-2.242	1.286	.050	.022			7.995	21.795	p=2
n=10	-10.628	9.323	.213	.172					p=1
n=7	-31.82	470.05	.596	8.546					p=1

Donde *se* es la desviación estándar. Aquellas celdas que figuran en blanco, es por que la variable no entra en el modelo.

Con el propósito de hacer una comparativa sobre la variación de los vectores beta, según numero de casos, medida en términos de estimación odd-ratio ajustado, se ejemplifica con los dos siguientes valores:

$$n=443 \quad \hat{odd}_{edad} = \exp(.0239) = 1.0242$$

$$n=7 \quad \hat{odd}_{edad} = \exp(.596) = 1.8148$$

3.6.4.2. Resultados algoritmo de Harrell.

Se incluyen en este epígrafe, a demás de los resultados resumidos en tablas, una ejecución completa del programa (Resultados 3.6.21.), cuyo fin es ilustrar el funcionamiento de este método.

Se ha aplicado el algoritmo de Harrell puro y corregido (véase epígrafes 2.9.6.3. y 2.9.6.4.) a los datos que claramente no cumplen la regla uno diez, esto es para n=10 y n=9 , tratando de conseguir con ello un Factor de Corrección o encogimiento (*shrinkage factor*). Se ha estimado el citado factor, por generación bootstrap, con al menos 300 generaciones. Se estiman índices pronóstico (*IP*) con vectores beta generados por bootstrap y aplicados a datos reales. En primer lugar y con el propósito de ilustrar el proceso de estimación, se editan algunas salidas intermedias del programa que implementa el algoritmo, para n=48, p=3 y B=300 generaciones bootstrap.

ESTIMACIÓN DEL MODELO ORIGINAL

Deviance = 3.9660e+01
Grados de libertad = 45.0

Estimación del vector beta y su error estándar

-4.9292	2.5782
0.0902	0.0425
9.4300	49.7619

GENERACION BOOTSTRAP 1. RESULTADOS.

Componentes del vector Beta, Original y Bootstrap, con su error estándar.

Compnte.: 0 Original: -4.929156 SE: 2.57816 Bootstrap: -7.604139 SE: 3.10135
Compnte.: 1 Original: 0.090209 SE: 0.04252 Bootstrap: 0.136988 SE: 0.05264
Compnte.: 2 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 3 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 4 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 5 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 6 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 7 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 8 Original: 9.429966 SE: 49.76191 Bootstrap: 9.284469 SE: 50.27320
Compnte.: 9 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000

Datos IP Reales y Bootstrap en fila 24 valen: 6.314611 y 9.589143

DEVIANCE's de los modelos Original, Bootstrap e IP: 39.6604 35.4074 56.9319

El Factor Corrector y su Error valen: FC: 0.001325 ERROR: 0.000000

GENERACION BOOTSTRAP 150. RESULTADOS.

Componentes del vector Beta, Original y Bootstrap, con su error estándar.

Compnte.: 0 Original: -4.929156 SE: 2.57816 Bootstrap: -7.498706 SE: 2.92669
Compnte.: 1 Original: 0.090209 SE: 0.04252 Bootstrap: 0.123255 SE: 0.04610
Compnte.: 2 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 3 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 4 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 5 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 6 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 7 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 8 Original: 9.429966 SE: 49.76191 Bootstrap: 8.459611 SE: 64.62777
Compnte.: 9 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000

Datos IP Reales y Bootstrap en fila 24 valen: 6.314611 y 8.627868

DEVIANCE's de los modelos Original, Bootstrap e IP: 39.6604 41.6674 63.7625

El Factor Corrector y su Error valen: FC: 0.001372 ERROR: 0.000000

GENERACION BOOTSTRAP 300. RESULTADOS.

Componentes del vector Beta, Original y Bootstrap, con su error estándar.

Compnte.: 0 Original: -4.929156 SE: 2.57816 Bootstrap: -2.749286 SE: 2.42876
Compnte.: 1 Original: 0.090209 SE: 0.04252 Bootstrap: 0.060179 SE: 0.04007
Compnte.: 2 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 3 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 4 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 5 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 6 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 7 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000
Compnte.: 8 Original: 9.429966 SE: 49.76191 Bootstrap: 8.966015 SE: 50.54956

Compnte.: 9 Original: 0.000000 SE: 0.00000 Bootstrap: 0.000000 SE: 0.00000

Datos IP Reales y Bootstrap en fila 24 valen: 6.314611 y 4.212520

DEVIANCE's de los modelos Original, Bootstrap e IP: 39.6604 38.3996 48.4522

El Factor Corrector y su Error valen: FC: 0.008777 ERROR: 0.000000

ALGUNOS ERRORES PARCIALES.

El FC en la generación bootstrap 0 vale: 0.001325
El FC en la generación bootstrap 3 vale: 0.000949
El FC en la generación bootstrap 4 vale: 0.006611
El FC en la generación bootstrap 150 vale: 0.001372
El FC en la generación bootstrap 154 vale: 0.003115
El FC en la generación bootstrap 155 vale: 0.012831
El FC en la generación bootstrap 296 vale: 0.012086
El FC en la generación bootstrap 297 vale: 0.000467
El FC en la generación bootstrap 298 vale: 0.002359
El FC en la generación bootstrap 299 vale: 0.002603

El Factor Corrector vale: 0.033667

El Factor Corrector, corregido, vale: 0.034005

Resultados 3.6.21.

La tabla 3.6.22 muestra los diversos Factores de Corrección, Harrel puro y corregido, según número de casos y variables incluidas en el modelo.

Tabla 3.6.22.

Número casos	FC (γ).	FC (corregido γ_c).	Número de vars.
n=86	.06175	.06846	p=2
n=48	.03367	.03400	p=2
n=10	.00316	.00352	p=1
n=7	.000764	.000819	p=1

Con el fin de apreciar los efectos generales del factor de corrección, sobre los vectores y odds-ratios, figura en Tabla 3.6.23., una comparativa de los vectores beta, corregidos por su correspondiente factor y las odds-ratios, corregidas y sin corregir.

Tabla 3.6.23. Comparativa de coeficientes y odd-ratios.

N° Casos. FC (γ_c)	β_0(Cte.)	β_1 (Edad)	β_8 (Diabetes)	Vars.
n=443 (Real)	-.3122	.0239	.9608	p=2
Odd-ratio		1.0241	2.6326	
n=86 (Corregido)	-2.242	.0342	0.5473	p=2
Odd-ratio Corregd./ Sin Corregir.		1.0348/1.0512	1.7286/2.966	
n=10 (Corregido)	-10.628	0.000756		p=1
Odd-ratio Corg./ SC.		1.00075/1.2378		
n=7 (Corregido)	-31.82	.000488		p=1
Odd-ratio Corg./ SC.		1.00048/1.8148		

Capítulo IV: Discusión y conclusiones.

4.1. Discusión.

Se propone en este epígrafe un análisis valorativo, en profundidad, de los resultados obtenidos en el capítulo 3. Se consideran varios sentidos valorativos: intrínseco de los propios valores, comparativamente con otros resultados y globalmente con otras variables; así como del propio método en sí mismo. Esta discusión pretende ir más allá de los aspectos cuantitativos, y, a través de los resultados, atisbar la vertiente cualitativa de las metodologías empleadas. Este capítulo aúna las doble vertientes biológica y estadística, contemplada es esta memoria como partes de un todo; evidenciando la complementariedad científica de ambas.

Se realiza un análisis previo, que comprende los dos tipos de variables contenidos en la base de datos: numéricas y categóricas. Este apartado utiliza diversas metodologías, que son objeto de valoración y comparación. Se analizan y discuten los diversos resultados de curvas ROC y análisis discriminante. Por último son objeto de valoración diversos resultados estimados mediante regresión logística.

Al principio del tercer capítulo se realiza una primera aproximación a los individuos contenidos en la base de datos. Se comentan las variables y se definen criterios biológicos de clasificación. Esta definición de variables y criterios se hace imprescindible, para entender todo el tratamiento posterior.

4.1.1. Análisis preliminar de potenciales factores numéricos y categóricos.

En Tabla 3.3.1. se recuentan, por sexos, y describen las poblaciones de nomotensos e hipertensos objeto del análisis. Destacar que las cifras medias o porcentuales que se refieren a hipertensos son mayores que las de normotensos: citar prevalencia de *diabéticos*, *obesidad*, *corazón* y *cerebro*; entre las numéricas destacar *triglicéridos y colesterol*. También ocurre que las prevalencias son generalmente, mayores en hombres que en mujeres. Los *p*-valores del *t-test* o de la *ji-cuadrado*, según variable numérica o categórica; referentes a la comparación por sexos, son en muchos casos significativos, evidencia claras diferencias sociosanitarias.

La discusión sobre factores numéricos y categóricos, que comprende las tablas y figuras de 3.2.2 a 3.2.9, ya se ha realizado en el tercer capítulo al exponer los resultados.

4.1.2. Estimación de curvas ROC.

Se presentan, en los epígrafes siguientes, el análisis de los resultados reflejados en el epígrafe 3.4, de la triple estimación de curvas ROC para el diagnóstico (Cruda, Normal y no Paramétrica). Se analiza cada una de las estimaciones tanto en sí misma, como comparativamente con las demás; prestando especial atención al área bajo la curva, que, en gran medida, evalúa la bondad del protocolo diagnóstico.

Para facilitar la correcta interpretación del segundo criterio cut-off, máxima Sensibilidad, cabe recordar que este criterio es básicamente igual al primero, máxima sensibilidad más especificidad; y además se exige, al menos un 90% de sensibilidad, si esta no se alcanzara, se tomará la cifra más próxima.

4.1.2.1. Análisis para la variable edad.

Considerando Resultados 3.4.1.; en la estimación cruda, el máximo de la sensibilidad más especificidad, se obtiene a los 65 años. En este pico no se alcanzan grandes valores, pues Sensibilidad=0.527 y Especificidad=0.744, el segundo valor , Especificidad, es más aceptable, pues descarta la enfermedad en casi el 75% de los individuos sanos. Sin embargo el protocolo solamente detecta la enfermedad aproximadamente en el 53% de los individuos afectados. El segundo criterio, máxima sensibilidad, se obtiene, como era de esperar, a una edad más temprana, 46 años, aumentando la sensibilidad a un 92,84% , sin embargo la contrapartida es un drástico descenso de la especificidad situándose en un 18.09%. El primer criterio proporciona un valor equilibrado entre Sensibilidad y Especificidad, mientras que el segundo tiene una gran Sensibilidad, a costa de un número importante de Falsos Positivos.

El intervalo proporcionado por el doble cut-off del los valores Predictivo Positivo y P. Negativo [19,65] aporta una probabilidad muy alta en el extremo inferior 0.9674, pero muy baja en el superior 0.3405.

Conviene señalar un dato fundamental, el área bajo la curva ROC vale 0.6217 y observando su representación gráfica, está cercana a bisectriz del primer cuadrante, este valor es bajo.

Respecto a la estimación basada en el supuesto de normalidad: analizada la supuesta normalidad de los datos de la variable edad mediante el *test K-S* se obtiene un *p*-valor=0.059, lo que significa una normalidad relativa, de ahí que los resultados , aún coincidiendo en lo fundamental con la estimación cruda, sean más inexactos; esto se ve reafirmado con el valor del área bajo la curva normal, 0.60634, inferior al de la estimación cruda. Una comparativa se encuentra en la figura 3.3.2.

Sobre la estimación no paramétrica hay que comentar la bondad de esta metodología, que proporciona valores muy parecidos a los de la estimación

cruda, pues la diferencia entre los cut-off de las dos estimaciones, para los diversos criterios, es mínima: la mayor de se registra en el extremo *VPN,* en la cruda es 19 y en la no paramétrica 18. Señalemos sobre este método, que proporciona unas estimaciones de probabilidad muy parecidas al crudo, y área bajo la curva igual al tercer decimal. Continuando con el análisis metodológico, comentemos lo siguiente sobre los valores de los parámetros de alisamiento. Como se dice en la 2.7.3., al exponer el diseño, cada bandwidth óptimo, h, debe cumplir: $h \to 0$; los valores obtenidos (Resultados 3.4.1.) son, respectivamente, 1.117 y 1.660; habida cuenta que los datos se miden en unidades enteras, si por la condición anterior fuesen menores que la unidad, no tendrían efecto de alisamiento alguno. Por lo tanto la condición se traduce en que han de ser mayores que la unidad, aunque como n=443, no deben ser mucho mayores que este valor. Evidentemente si n>>443, los valores de los parámetros podrían variar, pues seguirán cumpliendo la segunda condición del diseño: $nh \to \infty$.

Observando Resultados 3.4.1., es interesante señalar lo aproximado de los valores cut-off obtenidos para los distintos criterios, usando las tres metodologías. Por último digamos que la relación entre hipertensión y edad está documentada en los epígrafes 1.2.5.1 y 4.1.4.1.1. Hemos de resaltar que está constatado que la presión arterial sistólica aumenta con la edad de forma progresiva, desde la juventud hasta la vejez.

4.1.2.2. Análisis para la variable glucemia.

Observando Resultados 3.4.3, en la estimación cruda el primer cut-off, máximo de especificidad+sensibilidad, 101, sitúa la Sensibilidad=0.601 y la Especificidad=0.609, que resultan valores bajos para una prueba diagnóstica. Para el segundo cut-off (máx sensibilidad) se obtiene un valor de 86, con un valor muy alto de Sensibilidad=0.923, pero a cambio Especificidad tiene un valor irrelevante 0.164. El intervalo *VPN, VPP* [64,324] no proporciona resultados relevantes, pues la probabilidad correspondiente al extremo inferior es 1, pero esto es debido a que este valor, 64, es el mínimo de los procesados. La probabilidad para el extremo superior, 324, es 0.3042, resulta baja. El diagnóstico

de HTA, basado en el nivel de glucemia, es pobre, pues el área bajo la curva no llega al 60% . Una comparativa se encuentra en figura 3.4.4.

Ni siquiera mencionamos los resultados obtenidos en el supuesto de normalidad, pues el valor del área bajo la curva, 0.346, es irrelevante. Esto se debe a la falsedad del supuesto de normalidad (véase la figura 3.3.5 y la tabla 3.3.6)

En la estimación no paramétrica se vuelven a confirmar los buenos resultados proporcionados por esta metodología, pues los resultados de los cut-off son prácticamente coincidentes con los de la estimación cruda, excepto el *VPP*, que es donde más diferencia existe. Las áreas bajo las curvas, en las dos metodología consideradas, son prácticamente coincidentes ,0.599 y 0.601 respectivamente. Respecto a los parámetros de alisamiento, la situación es análoga a la expuesta al final del epígrafe 4.1.2.1.

4.1.3. Análisis discriminante.

Considerando la figura 3.5.1 no se distingue perfectamente la existencia de dos poblaciones, normotensos e hipertensos (control y caso). Como era de esperar, visto los resultados del epígrafe 3.4, existe un cierto grado de solapamiento. Obsérvese que ninguna de las dos variables tiene gran poder de discriminación, pues en las proyecciones sobre los ejes no se distinguen nítidamente los grupos de puntos. Por esta razón, recurrimos a un marcador conjunto proporcionado por la función canónica discriminante.

En esta función lineal *Z*, Tabla 3.4.2, es de destacar el gran peso del coeficiente $\beta_3 = 1.515$ debido a variable *diabetes,* no obstante conviene recordar que esta variable es dicotómica (0=ausencia, 1=presencia). La edad no tiene un coeficiente de gran peso $\beta_1 = 0.051$, pero es para cada año transcurrido.

La figura. 3.5.3 contiene la curva ROC para el marcador conjunto *Z*, en la que obtenemos un área bajo la curva de 0.706, lo que significa mediana eficacia del diagnóstico basado en estas variables. Con el propósito de completar este análisis, el lector debería de comparar este valor con los obtenidos en el epígrafe de curvas ROC, considerando las variable *edad* y *glucemia*, como predictores de HTA. (figuras 3.3.2 y 3.3.4.) cuyas áreas bajo la curva son 0.62 y 0.60 respectivamente. El aumento producido es apreciable.

4.1.3.1. Funciones de clasificación lineal de Fisher.

En Tabla 3.5.4 figuran las dos funciones de clasificación, una para cada clase; obsérvense los coeficiente correspondiente a la variable *diabetes* (-0.705 y 0.077 respectivamente); el de los normotensos es negativo, lo que disminuye el valor global del marcador conjunto *Z*, en términos contrarios nos referimos al coeficiente de los hipertensos. La variable *edad* con coeficientes 0.315 y 0.342 respectivamente, es de gran peso en esta maximización; compárense los coeficientes para normotensos e hipertensos, este último es mayor, lo que tendería a maximizar la función.

En Tabla 3.5.5. se resumen los resultados de la clasificación promovida por estas funciones en la base de datos considerada. Obsérvese que solo el 60.7% de los individuos están correctamente clasificados, lo que significa un resultado no muy bueno.

4.1.3.2. Selección de variables para discriminación.

En Tabla 3.5.6 figuran los resultados del algoritmo de selección de variables; observamos que las variables incluidas son las mismas que las usadas en los puntos anteriores. Contiene dos variables; *edad y diabetes* entran en la selección por este mismo orden y con scores cada vez más bajos. Es de destacar que de una pila de ocho posibles marcadores, solamente se seleccionan dos.

4.1.4. Regresión logística.

Previamente a la aplicación de la metodología, se ha realizado un análisis preliminar (crudo) de *hta* y factores de riesgo. Se trata de, aún considerando odd-ratios crudos, y sin descartar posibles factores de confusión, vislumbrar asociaciones, que posteriormente, en el análisis multidimensional, se verán confirmadas o desmentidas.

Se discuten los resultados obtenidos tras haber aplicado la metodología a tres variables diferentes. La primera considera la hipertensión como enfermedad y analiza los factores de riego de la misma. Al tratarse de un estudio de caso-control, el modelo no es predictivo, sin embargo nos estima las odd-ratios de las variables que explican la enfermedad; hecho que resulta capital en su tratamiento y prevención. También se aplica esta metodología a las variables *corazón* y *cerebro*, cuyos datos figuran en la base; en ambos casos se obtienen modelos predictivos, aunque nuestro principal interés es conocer los factores que intervienen en la enfermedad y sus odd-ratios. Se resume en la tabla 3.6.1. los tres modelos estimados, sus odd-ratios e intervalos de confianza.

Por último, señalar al lector que los resultados obtenidos, si bien tienen gran rigor científico, se debe de considerar que han sido elaborados a partir de una base de datos concreta; por ello, si bien puede ocurrir que los resultados sean consecuentes con lo científicamente aceptado, también puede pasar lo contrario. En todo caso esta memoria pretende ser una modesta contribución a la búsqueda de la verdad científica.

4.1.4.1. Modelo de RL para HTA.

Se trata de un estudio de caso-control, por lo tanto el modelo no es predictivo. Esto sucede porque, tanto el número de casos como de controles han sido definidos a priori por el investigador, y por lo tanto, sí se puede estimar la odds ratios para los dos factores de riesgo, con sus respectivos intervalos de confianza, pero no estima probabilidades de *hta* en función de los predictores; esta

última estimación solo se podría realizar si conociésemos la prevalencia de *hta* y, con ella, aplicásemos el teorema de Bayes (véase 2.3.1).

En Tabla 3.6.2. no se detectan factores de protección, ni parece que influyan variables como *sexo* o *hábitos sociales*. Esto podría deberse a que pudieran existir otras variables importantes no contempladas en la base de datos.

En el test bondad del ajuste de Hosmer y L., Tabla 3.6.3., observamos que el modelo no encaja muy bien (p<00.052), no obstante tampoco se rechaza. Este hecho resultaría importante si se tratara de un modelo de carácter predictivo, pero no es nuestro caso.

Para valores de influencia o con residuales largos, Tabla 3.6.4., solamente cabe destacar el control, número 394, con un residual normalizado de –3.306 y un valor de influencia h=-1.188, que corresponde, como era de esperar, a una persona mayor (73 años) y diabética. En Resultados 3.6.5., se recogen las variaciones en el modelo registradas como consecuencia de la eliminación del mencionado individuo (394). Las variables en la ecuación son las mismas, solamente varía, al alza, la odd-ratio de la diabetes, que pasa a valer 2.979, IC (95%) [1.429, 6.210], cuando antes de la eliminación su valor era 2.683 , IC (95%) [1.326, 5.430]. El nuevo modelo es más rico en información, pues su AIC es 456.200, mientras que sin el individuo eliminado es mayor, 461.259. Origina residuos menores que el anterior, pues el máximo valor es < 3.1. Sin embargo, el test de HyL da un p-valor significativo (p=0.017), con lo que el modelo se rechaza. Luego, aún siendo este modelo más elaborado que el anterior, es preferible el primero.

Las variables contenidas en el modelo, son *edad y diabetes*, que se discutirán en los dos epígrafes siguientes. No obstante, previamente, conviene hacer algunos comentarios sobre el papel de los *lípidos* en la hipertensión arterial. Si bien es cierto que ninguna variable que represente a los lípidos (*colesterol y triglicéridos*) está en el modelo, sin embargo en el epígrafe 3.3, Tabla 3.3.7., ya se dejó constancia de que las medias de los variables *colesterol, triglicéridos*, en

la cohorte de los hipertensos, eran superiores en los de cohorte de la normotensos. También, en la misma tabla, considerando la variable *triglicéridos*, existe una clara evidencia ($p=0.026$), de diferencia significativa entre las medias (cualquiera que sean estas) de casos y controles. Por lo tanto, aún no estando estrictamente incluida en el modelo, si parece propio analizar el papel de dislipemia en la hipertensión arterial.

Varios estudios epidemiológicos han demostrado la mayor prevalencia de alteraciones lipídicas en los individuos hipertensos en relación con los que tienen una presión arterial dentro de límites normales. Se ha encontrado una correlación positiva y consistente entre ambos factores, lo que sugiere una interrelación continua, como justifica el estudio TRÖMSO. Otras estudios más recientes también han confirmado dicha relación, llegando a encontrarse entre pacientes hipertensos hasta un 41% de hipercolesterolemia y un 50% de hipertrigliceridemia.

Si el análisis epidemiológico se realiza desde el punto de vista de los individuos dislipémicos, también se observa que la presión arterial resulta superior en los hipercolesterolémicos que en los que presentan cifras normales de colesterol y que la relación entre colesterolemia y presión arterial es mayor para la presión arterial sistólica que para la diastólica. También se ha demostrado que la prevalencia de hijos dislipémicos de padres hipertensos, es mayor que en la población general. Los individuos hipertensos muestran concentraciones plasmáticas de colesterol total y de colesterol LDL superiores a los de la población normotensa. Recientemente se ha descrito que los sujetos con hiperlipemia familiar combinada y aumento de la grasa visceral, presentan una mayor actividad simpática debido al aumento del flujo de ácidos grasos al hígado, lo que contribuye fuertemente a la elevada prevalencia de la denominada hipertensión dislipémica familiar.

En modelos experimentales se ha observado que la penetración de las LDL en la pared arterial y su papel en la génesis de la arteriosclerosis, no depende

exclusivamente de la concentración de estas partículas en el medio. Sino también de la presión que se induce en el modelo. El estrés hemodinámico juega un papel clave en el desarrollo de lesiones vasculares como revela el hecho de que las placas de ateroma predominen en los puntos de bifurcación arterial, que las venas sólo las presenten en los injertos arteriolizados, y que en las condiciones fisiológicas de baja presión de las arterias pulmonares no se encuentran lesiones arterioscleróticas, que sí se desarrollan en los pacientes con hipertensión pulmonar.

Por otro lado, la elevación mantenida de las cifras de colesterol puede favorecer las respuestas vasoconstrictoras periféricas, ya que disminuye la producción de oxido nítrico endotelial y también aumenta su catabolismo, debido a la oxidación lipídica sobre las placas de ateroma.

También se sabe que las lipoproteínas oxidadas activan la proteincinasa en la membrana celular y como consecuencia ocurre la activación de contratransporte Na/H+ con aumento de la entrada de sodio y el aumento del calcio en el citosol libre, lo que propicia la vasoconstricción en las células del músculo liso arteriolar. Por tanto, todo apunta a que la asociación hipertensión arterial - dislipemia no es una mera coincidencia, ya que se perfilan nexos patogénicos entre ambos factores de riesgo cardiovascular.

En los dos siguientes epígrafes se discuten cada una de las variable contenidas en el modelo.

4.1.4.1.1. Variables en el modelo: Edad.

En Tabla 3.6.2., observamos que la primera variable que entra en el modelo es la edad; con una odd-ratio de 1.024 y un IC al 95% [1.007, 1.041], si bien la odd-ratio pudiera parecer baja, debe contemplarse que es por cada año transcurrido. El IC es estrecho y su extremo inferior está por encima de la unidad.

La hipertensión arterial es un problema frecuente en las personas de edad. Se sabe que la presión arterial, tanto la sistólica como la diastólica, aumentan con la edad. El estudio Framingham también demostró que el riesgo cardiovascular no declina con la edad, tanto en el caso de la hipertensión diastólica, como la sistólica aislada.

La prevalencia de la hipertensión arterial además se ha multiplicado tanto por el efecto del envejecimiento de la población y el aumento de la expectativa de vida, como por la modificación de los criterios diagnósticos, que recientemente se han fijado y aceptado, en las cifras límite de 140 para la presión arterial sistólica, y en 90, para la diastólica. Estos límites han venido a sustituir los anteriormente propugnados por la OMS, de 160/95. Ya en su informe de 1999, la OMS junto con la Sociedad Internacional de Hipertensión, se ha adherido a aquellos mismo criterios.

Está constatado que , salvo en comunidades de vida muy primitiva, la presión arterial sistólica aumenta con la edad de forma progresiva, desde la juventud hasta la vejez. Este hecho es más evidente en el sexo femenino. La presión arterial diastólica, por el contrario, tiende a mantenerse o incluso a disminuir a partir de la sexta década de la vida.
La presión arterial es el resultado de la relación entre el gasto cardiaco y las resistencias periféricas. Otras variables modulan las respuestas a ambos factores: actividad del sistema nervioso simpático, sistema renina-angiotensina, reflejos barorreceptores, factores natriuréticos circulantes, y el endotelio vascular, a través de la secreción de sustancias vasoactivas como el óxido nítrico, las endotelinas y las prostaciclinas.

Conforme avanza la edad, y especialmente en el anciano, a la caída del gasto sistólico y a la menor sensibilidad de los basorreceptores, como elementos diferenciadores, se suma la disminución de la *compliance* de los grandes vasos, que junto con la elevación de las resistencias periféricas, origina una elevación de las cifras de presión arterial sistólica. Esto explica que en la clínica diaria, la forma más habitual de hipertensión en el anciano sea la hipertensión arterial sistólica , y que conforme aumenta la edad de la población estudiada, la prevalencia de este tipo de hipertensión sea cada vez más elevada y alcance tasas de hasta un 70% en los tramos de mayor edad. No siempre ha sido así, porque hasta hace pocos años se consideraba a la hipertensión sistólica aislada, como un "factor de compensación" ante el aumento de las resistencias periféricas. Pero hoy en día existen suficientes evidencias de que la hipertensión arterial aislada no solo no es un "factor de compensación", sino que constituye un factor de riesgo de morbimortalidad cardiovascular, incluso con valor predictivo superior que la hipertensión diastólica.

La relación entre hipertensión y edad se ve reforzada por las evidencias epidemiológicas. La prevalencia de HTA en los mayores de 65 años esta estimada en un 70%, frente al 34.2% de prevalencia establecida para la población adulta situada entre los 35 y los 64 años.

4.1.4.1.2. Variables en el modelo: Diabetes.

Como figura en la tabla 3.6.2., la otra variable que entra en el modelo es diabetes, que claramente es un factor de riesgo con una odd-ratio alto, 2.683, y un Intervalo de Confianza al 95% [1.326,5.430]; pues el extremo inferior es claramente superior a la unidad. Dicho de otro modo: la fracción de hipertensos en la cohorte de diabéticos es 2.683 superior a los existentes en la cohorte de los no diabéticos, y nos informa del riesgo que supone la diabetes para ser hipertenso.

La elevación de los valores de la presión arterial en los diabéticos tiene implicaciones pronósticas, que han sido puestas en evidencia en los últimos años. En la diabetes, la hipertensión se acompaña de un incremento de la morbimortalidad, como consecuencia de un mayor riesgo de complicaciones tanto microvasculares, como la nefropatía y la retinopatía; como macrovasculares: cardiopatía isquémica, insuficiencia cardiaca congestiva y arteriopatía periférica.

Las complicaciones micro y macrovasculares, se desarrollan en paralelo, pero en el caso de la diabetes tipo uno, la edad temprana de comienzo de la enfermedad, condiciona que la nefropatía adquiera un protagonismo mayor en las primeras décadas de la enfermedad. La presencia de la nefropatía diabética acelera el desarrollo de complicaciones cardiovasculares, en parte por la elevación de las cifras de presión arterial que la acompañan. Se sabe que el tratamiento antihipertensivo enlentece el desarrollo de estas lesiones, reduciendo la incidencia de insuficiencia renal terminal y de accidentes cardiovasculares.

La nefropatía diabética se desarrolla aproximadamente en el 25-30% de los diabéticos tipo uno, y constituye una de las complicaciones clave que determinan la morbimortalidad y la expectativa de vida.

La gravedad de la hipertensión en la diabetes tipo uno, sigue paralela al desarrollo de la nefropatía. Mientras que en los sujetos sin nefropatía, la prevalencia de HTA es similar a la población general de la misma edad y sexo, en los que sufren nefropatía incipiente la prevalencia se eleva ligeramente. En los sujetos con nefropatía establecida, la

prevalencia de HTA es superior al 60% y se incrementa al 90% cuando la nefropatía se grava y el filtrado glomerular es inferior al 30 ml/min.
También está establecido que la reducción de los valores de presión arterial con tratamiento antihipertensivo retrasa la velocidad de progresión de la nefropatía en una proporción que depende de los valores de presión arterial alcanzados.

Estudios más recientes también han demostrado la efectividad del tratamiento antihipertensivo en pacientes con diabetes tipo dos e hipertensión. Así el estudio SHEP (Sistolic Hipertension in the Elderly Program), en mayores de 60 años, con diabetes e hipertensión, demostró una reducción de los eventos cardiovasculares en un 34% en los que seguían tratamiento antihipertensivo, frente a los que solo recibían placebo.

También en el Sist-Eur (Systolic Hypertension in Europe) realizado en mayores de 60 años con hipertensión sistólica aislada, se demostró que en el subgrupo de los diabéticos tipo dos, se producía una reducción del números de accidentes cerobrovasculares fatales en un 42%, y de eventos cardiovasculares en un 26%; al disminuir la presión sistólica con tratamiento.

4.1.4.2. Modelo de RL para corazón.

Vistos los grandes residuos, Tabla 3.6.9, que genera el primer modelo, con el fin de mejorarlo, resulta interesante eliminar algunos. El modelo así generado es más rico en información que el primer modelo; pues su ACI final es, Tabla 3.6.12., 465.117, inferior a la del modelo anterior, Tabla 3.6.7, 470.221. El encaje de los modelos es bueno y prácticamente el mismo; el primero Tabla 3.6.8, (p=0.70), y el segundo o depurado, Tabla 3.6.11., (p=0.69). Comparando Tablas 3.6.9 y 3.6.14, observamos que los residuos generados en el segundo modelo son

más cortos que los del primer modelo (sin depurar) , pues en el segundo apenas si sobrepasan el valor 3 veces la desviación típica.

Considerando Tabla 3.6.10 observamos que los factores que intervienen en las dolencias cardiacas son varios y de diversa índole. La *edad* se encuentra presente con una odd- ratio de 1.035 IC [1.016, 1.056] , por cada año transcurrido. Por otro lado, *sexo* odd-ratio 2.255 IC [1.423, 3.575], *diabetes* odd- ratio 2.151 IC [1.299, 3.564] son factores de riesgo importantes. La variable *hta* odd-ratio 2.567 IC [1.250, 5.272] es, aisladamente, el mayor factor de riesgo. Ninguno de los intervalos de confianza de las variables en la ecuación contiene a 1, por lo tanto todos son verdaderos factores de riesgo. Las odds-ratios están más pulidas y el modelo depurado es , en general, mejor que el primero considerado.

Como el modelo es predictivo se incluyen la tabla de clasificación 3.6.13, que no es otra cosa que una tabla de contingencia, donde figuran enfermos y sanos observados, y pronosticados por el modelo, en filas y columnas respectivamente. El modelo clasifica correctamente el 74.5% de los individuos, lo que supone un alto porcentaje.

Por último comentamos figura 3.6.15 (a) y (b); en ella se representa el par $(\hat{\pi}, \Delta X^2)$, que es una de las principales mediadas de dispersión de modelo. Estos dos diagramas se encuentran en relación con las tablas 3.6.9 y 3.6.14 respectivamente. Los puntos que presentan un pobre encaje son los situados, en la parte más alta de las graficas, en los extremos derecho e izquierdo Así la gráfica 3.6.15.(a) (Modelo no depurado) presenta, para el patrón Y=0, un punto por encima de 5, que es claramente un outlier; para Y=1, hay 3 puntos por encima del 5. Del mismo modo que en Tabla 3.6.14. los residuos son menores que en Tabla 3.6.9, en la gráfica 3.6.15. (b) (Modelo depurado) los puntos de peor encaje han desaparecido: para Y=0 el único outlier ya no está. Para el patrón Y=1, de los tres puntos con ΔX^2>5, tan solo queda uno, y es la abcisa más pequeña. Queda patente que el Modelo depurado tiene una menor dispersión. No obstante, al depurar, según se refleja en la gráfica 3.6.15.(b), también se pierde información,

pues en la segunda grafica no existen puntos con probabilidades predichas mayor que 0.7, cuando, observando la gráfica 3.6.15.(a), existen cuatro con esta característica.

A continuación se considera la vertiente biológica de los resultados, realizándose diversas consideraciones sobre los mismos.La causa más frecuente de morbimortalidad en los pacientes hipertensos la constituye la cardiopatía hipertensiva. Se conoce con este nombre al conjunto de alteraciones estructurales y funcionales del corazón relacionadas con la hipertensión arterial.

La primera manifestación objetivable, de alteración estructural, que puede apreciarse en el tiempo, sobre el corazón del individuo hipertenso, es la hipertrofia ventricular izquierda. En el desarrollo de la misma, además de la propia hipertensión arterial, influyen otros factores como edad, raza, sexo, peso, composición corporal y actividad de diversos sistemas tróficos: sistema adrenégico y sistema renina - angiotensina – aldosterona.

El miocardio hipertrófico se caracteriza por un aumento del volumen de los miocitos, con cambios en la composición de las proteínas contráctiles, mayor volumen fibrótico y disfunción de la microcirculación coronaria, que junto con las lesiones arterioscleróticas y una posible compresión de las arterias intramiocárdicas por un miocardio hipertrófico y fibrótico, dan lugar a una reducción de la reserva coronaria. Dichas modificaciones establecen el nexo patogénico con la disfunción ventricular, la isquemia miocárdica y las arritmias. Las alteraciones de la hemodinámica coronaria podrían explicar el incremento de la prevalencia

de arritmias cardiacas, isquemia miocárdica, sobre todo silente y de muerte súbita, asociadas a la hipertrofia ventricular izquierda.

La hipertrofia ventricular izquierda hipertensiva, es un factor de riesgo cardiovascular de primer orden. Ello ha sido demostrado por diversos estudios prospectivos. La prevalencia de esta entidad aumenta con la edad. Utilizando la ecocardiografía, se encontró un incremento de su prevalencia desde el 8% en varones menores de 30 años, hasta un 33% en mayores de 70.

Un estudio llevado a cabo por Coca y colaboradores- estudio VITAE- analizó la prevalencia de hipereterofia ventricular izquierda en población hipertensa, encontró unas cifras de 61% en varones y un 46%en mujeres en el grupo de edad situado entre 30 y 64 años, mientras que para los mayores de 65 años, la prevalencia de hipertrofia ventricular izquierda se situó en 73% para los varones y en 63% para la mujeres.

Otra complicación destacada de la cardiopatía hipertensiva, la constituye la isquemia miocárdica. En su origen se encuentra la aterosclerosis coronaria, la enfermedad microvascular o una combinación de ambos procesos. Existe una relación epidemiológica demostrada por varios estudios entre cardiopatía isquémica (angina estable e inestable e infarto agudo de miocardio) e hipertensión arterial. Un estudio reciente observó que cerca de la mitad de los pacientes ingresados por infarto agudo de miocardio se conocían hipertensos.

La hipertensión arterial también representa un factor etiológico de considerable importancia en la insuficiencia cardiaca. En el estudio CARDOTENS 99, el 77% de los pacientes diagnosticados de esta

patología, lo estaban también de hipertensión arterial. El paso de hipertensión arterial a insuficiencia cardiaca por disfunción sistólica es un proceso de patogenia multifactorial, cuya fisiopatología no está del todo aclarada, que afecta a todos los componentes celulares miocárdicas.
La hipertensión arterial también es el factor más frecuentemente asociado a la fibrilación auricular crónica, que se observa en el 66% de los pacientes hipertensos.

El conjunto de las alteraciones funcionales y estructurales del corazón que se observan en la hipertensión arterial, constituyen el determinante fundamental asociado a la mortalidad de los pacientes con hipertensión. El control estricto de la presión arterial en estos pacientes, constituye la principal estrategia contra la evolución fatal de la enfermedad hipertensiva.

4.1.4.3. Modelo de RL para cerebro.

La enfermedad cerebrovascular, accidente cerebrovascular o ictus, constituye una patología de elevada incidencia y mortalidad, y ocasiona situaciones muy frecuentes de incapacidad en diversos grados. Se calcula que la incidencia se encuentra entre los 150 y 200 casos por cien mil habitantes y año, y que la prevalencia media es de 600 caso por cien mil habitantes. En los últimos años se observa un descenso de la incidencia, debido al mejor control de los factores de riesgo.

La enfermedad cerebrovascular aguda se considera la tercera causa de mortalidad, después de la enfermedad cardiovascular y el cáncer y ocasiona entre le 12 y el 15% de las muertes anuales. De estas el 88% ocurren en personas mayores de 65 años. La mortalidad se encuentra

alrededor del 20% en la fase aguda de la enfermedad y se eleva al 25% si se considera el primer año tras el episodio.

De los pacientes que sobreviven al primer mes, cerca del 50% presentarán algún grado de incapacidad que genera dependencia. En un 30% de los casos dicha incapacidad puede considerarse severa.
En las últimas décadas la mortalidad de la enfermedad cerebrovascular aguda se ha reducido en un 7%, gracias entre otros al mejor control de factores de riesgo como la hipertensión.

Los factores de riesgo que intervienen en estas patologías , Tabla 3.6.17, son varios y de distinta índole. La *edad*, en sentido de envejecimiento, se encuentra presente con una odd- ratio de 1.078 IC [1.034, 1.123] , por cada año transcurrido. El *sexo* también está en el modelo, odd-ratio 4.996 IC [1.979, 12.612]. La *diabetes*, odd-ratio 5.334 IC [1.979, 12.612], es un factor de riesgo muy importante. Desde el punto de vista biológico, las odds-ratios de *sexo* y *diabetes*, pueden resultar demasiado grandes. Ninguno de los intervalos de confianza de las variables en la ecuación contiene a la unidad, por lo tanto todos son verdaderos factores de riesgo.

La variable *hta* no entra en el modelo, sin embargo entre los factores de riesgo, el sustrato anatomopatológico más frecuentemente encontrado en el origen del ictus, es la arteriosclerosis. La hipertensión arterial es el factor de riesgo modificable más potente, tanto en lo que se refiere a la hipertensión arterial sistólica, como a la diastólica. La hipertensión arterial sistólica, tiene incluso un valor predictivo mayor, y es independiente de la edad.

La información del modelo ACI, Tabla 3.6.16., es 182.896, que resulta un valor bajo. En el test bondad del ajuste de HyL, Tabla 3.6.18., observamos que el

modelo encaja bien (p=0.23). Por último, en Resultado 3.5.19., se expone el modelo predictivo de un modo explícito.

4.1.4.4. Corrección de la sobreestimación del vector beta de coeficientes de RL.

En los siguientes epígrafes se discuten los resultados obtenidos por generación bootstrap, en referencia a inflación o sobrevaloración de los coeficientes del vector beta en relación con el número de variables y casos. También se analizan los distintos factores de corrección, así como su influencia sobre el citado vector. Es de considerar la simulación realizada en el epígrafe 2.9.6.5., pues abarca un importante segmento de potencialidades, mientras que la aplicación a la base de datos supone el contraste con la realidad.

Analizando el vector beta de coeficientes, considerando Tabla 3.6.20, es de destacar el hecho de que disminuyendo el número de casos hay variables que no entran en el modelo; aún más, según se restringen los casos, también disminuye el número de variables presente en el modelo. Para ilustrar este hecho basta comprobar que para n=443, además de la constante, hay dos variables (*edad* y *diabetes*) en el modelo (p=2) y por último para n=10 la variable *diabetes* ya no se halla presente en el modelo (p=1). No obstante, este hecho está suficientemente claro en la simulación del epígrafe 2.9.6.5.

Por otro lado, resulta importante observar la inflación o aumento de los coeficientes β_i, $\forall i = 1,\cdots,p.$ de las variables incluidas en el modelo, según disminuye el número de casos. Esto sucede en cualquier variación, a la baja , del número de datos. Obsérvese la gran diferencia para n=443 entre β_1= .0239 (edad), y el mismo coeficiente β_1=.596 para n=7. Esta diferencia, medida en términos de odd-ratio, resulta importante: $\hat{odd}_{edad} = \exp(.0234) = 1.0242$ y $\hat{odd}_{edad} = \exp(.596) = 1.8148$.

En la misma línea, resulta interesante observar el aumento registrado en los errores estándar (*se*) a medida que disminuye el número de casos. Considérese,

a modo de ejemplo, *edad*, su valor en claro aumento según disminuye el número de casos; llegando a tomar un valor muy grande: $\beta_1 = .596$ y *se*=8.546.

Es importante evaluar el número de datos y sus repercusiones en el modelo, reacuérdese la regla 1-10. Este hecho está ilustrado en Tabla 3.6.20, Obsérvese que cuando se utilizan un gran volumen *n*=443, están en el modelo *edad y diabetes*, cuyos valores son: *edad* .0239 y *diabetes* .9608. Son de señalar las grandes variaciones, al alza, registradas en los coeficientes del vector beta cuando existen pocos datos, y se reducen. Considérese la variable edad, que está siempre en el modelo, para *n*=10, $\beta_1 = .213$, con solo tres datos menos *n*=7 su valor aumenta a $\beta_1 = .596$.

4.1.4.4.1. Análisis de los factores de corrección.

Como pauta general, es preciso destacar que los factores de corrección corren paralelos al vector beta; esto es, cuantos menos datos contiene el modelo, mayor es el sobredimensionamiento del vector beta, y más ha de "encogerse" el mismo; por lo tanto, más pequeño será el Factor de Corrección.

En Tabla 3.6.22 se muestra la evolución de los factores de corrección según disminuye el número de casos, cuyos resultados son del todo concordantes con los de la simulación, aunque aquí se trata una escasez de datos extrema. Obsérvese como según esto ocurre, paralelamente el FC se hace más pequeño, así para *n* =86 $\gamma = .06175$ y para *n*=10 disminuye hasta $\gamma = .00316$. También es de destacar, cuando los casos son muy pocos, la importante disminución que se produce eliminando algunos, así para *n*=10 es $\gamma = .00316$, mientras que para *n*=7 el factor disminuye hasta $\gamma = .000764$. Previamente, en Resultados 3.6.21, se da la salida del programa, que ilustra por pasos la estimación de los factores de corrección.

En la misma tabla 3.6.22 figura el FC corregido ,γ_c, cuya finalidad es suavizar el FC ,γ, ya que en la práctica se demuestra que este valor aproxima

mejor el vector beta a su valor real, ya que el FC puro lo encoge demasiado. El FC corregido no es más que el resultado de eliminar el 1% de los valores más pequeños, de los generados mediante el bootstrap y usados para calcular la media que nos da el FC, de ahí que sus valores sean ligeramente superiores a los del FC puro.

La Tabla 3.6.22. contiene una comparativa de los vectores beta real y corregido, así como de las dos odds-ratios, la proveniente del vector beta corregido y sin corregir. Como podemos observar, el vector beta corregido está más cerca del real que el sin corregir. Las odds-ratios, provenientes del beta corregido, toman valores más próximos a las reales.

4.2. Conclusiones.

Se exponen, en los epígrafes siguientes, las conclusiones obtenidas después de aplicar las diversas metodologías expuestas en esta memoria. No se pretende realizar un metaanálisis, mas bien se trata de concluir con rigurosidad y cuidando el aspecto didáctico, sobre las herramientas utilizadas y los resultados obtenidos.

4.2.1. Análisis preliminares.

El recuento, desglose y primera aproximación al conocimiento de las poblaciones, resulta imprescindible. El análisis previo de datos numéricos y categóricos se revela como una herramienta eficaz para detectar factores de confusión. Así el ajuste de medias nos informa sobre las ligeras perturbaciones causadas por la variable *edad* sobre las otras cuatro variables numéricas. También muestra indicios sobre posibles asociaciones e influencias. Resulta importante,

ser extremadamente cuidadoso con la elección de parámetros de resumen de la población.

Los resultados del test de las permutaciones bootstrap ponen de manifiesto que esta método proporciona resultados muy exactos con muy pocas generaciones. El bootstrap, como metodología general, ha sido ampliamente usado en esta memoria; la bondad de las aproximaciones realizadas está fuera de toda duda; véase, entre otros ejemplos, Tabla 2.6.4. Se valida como un método de grandes posibilidades.

En términos parecidos a lo dicho sobre el análisis previo de variables numéricas, nos podríamos expresar sobre el análisis preliminar de variables categóricas. Este tipo de análisis, un primer borrador, induce cuestiones, que sugieren el camino a seguir.

4.2.1.1. Transformaciones de Box y Cox.

Estas transformaciones siempre logran reducir, de un modo muy significativo (mínimo el 20%), la dispersión de los datos. Las puntuaciones extremas (outliers) pueden dejar de serlo, pero en todo caso se encuentran más centradas y, en general, todas las puntuaciones se hacen más compactas, disminuyendo de un modo muy claro la dispersión.

Mejoran las medidas de forma, asimetría y curtosis, tomando la distribución una forma más parecida a una normal que aquella que tenía anteriormente.

Hemos de resaltar los valores que toma la potencia en las distintas transformaciones, en general muy próximos a cero; cuestión que no depende del rango de los datos, por lo tanto cabe concluir que generalmente las mejores transformaciones se encuentran cercanas a la logarítmica, e incluso pueden coincidir con esta función. Aunque esto no constituye una verdad absoluta.

La normalidad, objetivo prioritario de las transformaciones de ByC, no se consigue siempre conjuntamente para casos y controles. Aquí, y a la vista de la discusión anterior, concluimos que si los datos no están muy distanciados de la hipótesis de normalidad (variable *triglicéridos*), esta se consigue conjuntamente. Por el contrario si están muy alejados (basta con que lo esté uno de los dos grupos caso o control, variables *glucemia y ácido úrico*) los datos transformados tampoco presentan normalidad.

Resulta redundante aplicar la transformación cuando ya los datos presentan normalidad (variable *colesterol* y *edad*), pues si bien se mejora la dispersión, se empeora la forma y se pierde la normalidad.

4.2.2. Estimación de curvas ROC.

La curva ROC proporciona, en términos de área bajo la curva, un método fiable para evaluar la bondad de un protocolo diagnóstico. Esta afirmación viene avalada por, además de la simulación, los resultado obtenidos para las dos variables, *edad y glucemia*, indicadores para el diagnóstico de HTA. En ningún caso sobrepasan un área del 63% , lo que nos dice que son indicadores, pero no bastan cada uno por sí solo para obtener un buen diagnóstico; este se obtendría, al menos, con más de un 80% de área bajo la curva. Obsérvese que en el modelo de *hta* (epígrafe 4.1.4.1.), existen más variables predictoras además de *edad.* También se incluye *diabetes*, relacionada con *glucemia*, pero ninguna de ellas, especialmente *edad*, que proporciona mejores resultados, es un buen marcador, por si solo, para un diagnóstico de HTA.

Los tres criterios que responden a diferentes necesidades de conocimiento, bien sea de equilibrio entre detectar y rechazar la enfermedad, diagnóstico fiable, o un intervalo para pruebas complementarias, cumplen su objetivo.

La estimación normal resulta valiosa siempre y cuando los datos sigan esta distribución.

Los resultados proporcionados por la estimación no paramétrica son bastante aceptables, tanto en términos de área bajo la curva (los valores que toman son prácticamente los mismos), como en puntos cut-off y sus probabilidades, cuyas puntuaciones son muy aproximadas. Su función es proporcionar una orientación fiable sobre los resultados reales.

Se valida el método de estimación no paramétrica, usando estimadores de núcleo y se contrasta la bondad de las metodología de curvas ROC, cuyos resultados son acordes con los obtenidos usando regresión logística.

4.2.3. Análisis discriminante.

El marcador conjunto estimado por este método, a pesar de ser de mediana eficiencia, supera en términos de área bajo la curva ROC (0.706 frente a 0.62), los resultados obtenidos con los marcadores individualizados *edad* y *glucemia.*

La selección de variables es acorde con los resultados obtenidos en regresión logística, pues contiene únicamente *edad* y *diabetes,* únicos factores que forman parte del modelo de regresión logística.

La clasificación promovida por las funciones de Fisher para asignar individuos a grupos, según su vector de datos, en este caso no es demasiado eficaz, aunque las potencialidades que encierra lo hacen de interesante aplicación en ciencias de la vida.

4.2.4. Modelos de regresión logística.

Las conclusiones, para los modelos de RL, las expondremos desde un doble punto de vista; el primero abarca las consideraciones sobre el modelo propiamente dicho: odd-ratios, encaje, residuales u otros aspectos, que resulten de interés; este punto se desarrolla en el epígrafe 4.2.4.1. El segundo se refiere a las factores de riesgo para HTA, o enfermedades relacionadas, que están en los diferentes modelos. Este segundo punto, se desarrolla en 4.2.4.2.

4.2.4.1. Modelos de RL per se.

El principal interés de los modelos de RL, se centra en las variables que contiene, el ajuste por la variable *edad,* y la estimación de las odd-ratios correspondientes a los factores de riesgo o protección. Para avalar este hecho, basta constatar que ningún IC para las odd-ratios, contiene a la unidad, con lo que todas las variables incluidas son verdaderos factores de riesgo. Observamos, en la construcción de los tres modelos, como van desapareciendo los factores de confusión y de las múltiples variables candidatas a entrar en el modelo, finalmente solo están unas pocas. Queda clara la validez de los modelos de RL, y, por ende, del análisis multivariante.

El hecho de que el modelo sea predictivo, como es el caso de las variables *corazón y cerebro*; o no predictivo, como en el caso de la variable *hta*; no es de fundamental importancia, pues en el primer caso podemos estimar la probabilidad de padecer la enfermedad y obtener una tabla de clasificación, como es el caso de la variable *corazón*; nuestra idea sobre la enfermedad no es predictiva, sino preventiva y terapéutica; por ello, la cuestión realmente importante es conocer los factores de riesgo y sus odd-ratios.

Los diferentes índices que estimamos en los modelos, tales como ACI o Residuos, tiene una función valorativa del mismo; así el ACI siempre evoluciona a la baja. El modelo final es el más depurado, y objeto de mayor valoración, siempre su ACI es más pequeña que la de los anteriores, y si origina grandes

residuos, como en le caso de la variable corazón, es conveniente eliminar estos casos y desarrollar un nuevo modelo. El test de HyL valora la bondad o encaje del modelo. Todos lo modelos estimados en esta memoria tienen un encaje aceptable.

4.2.4.2. Hipertensión arterial y factores de riesgo.

4.2.4.2.1. Hipertensión arterial, Edad y Sexo.

En los distintos modelos estimados, específicos para *hta* o relacionados, la *edad* se encuentra siempre presente, con una odd-ratio *mínima* de 1.024, por cada año transcurrido. Quizás sea la *edad* el factor no modificable más importante que se encuentra asociado a la hipertensión. Está perfectamente documentado que las cifras de tensión arterial aumentan con la *edad.* Durante la infancia y la adolescencia los principales determinantes del incremento de la tensión, son los cambios ligados al crecimiento: aumento de peso y talla. Estos dos factores son responsables de un aumento anual de aproximadamente 1.5 mm Hg para la tensión arterial sistólica y de 0.5 mm Hg para la sistólica, entre los seis y los 18 años.

Posteriormente, las cifras van aumentando progresivamente a lo largo de la vida en ambos sexos, pero siempre más bajas en las mujeres que en los hombres, hasta la menopausia, momento en que la TA de la mujer aumenta muy rápidamente y llega a sobrepasar los valores promedios del varón. No obstante, mientras que la presión diastólica se estabiliza, e incluso desciende a partir de los cincuenta años, la sistólica sigue aumentando en ambos sexos, lo que justifica la alta prevalencia de HTA y de HTA sistólica aislada en los sujetos de más de 75 años.

La mayor esperanza de vida de la población occidental ligada al descenso de las tasas de mortalidad y al bajo índice de natalidad, conducen al progresivo envejecimiento poblacional, lo que se acompaña de un aumento de la

prevalencia de enfermedades crónicas y degenerativas asociadas a edades avanzadas, entre las que destaca la HTA. Los hábitos higiénicos y dietéticos de las sociedades desarrolladas favorecen la progresión de la arteriosclerosis y determinan un progresivo aumento de la rigidez de las grandes arterias elásticas, lo que se suma a fenómenos propios del envejecimiento como la sustitución de tejido elástico por colágeno en vasos y corazón.

La HTA afecta a más de un 25% de la población. Pero la prevalencia aumenta con la edad de modo que por encima de los sesenta y cinco hasta el cincuenta por ciento de la población presenta cifras de presión sistólica superiores a 140/90 mm Hg, o cifras de presión sistólica superiores a 140 mm Hg con diastólica inferior a 90mm Hg, entidad conocida como HTA sistólica.

El hipertenso anciano tiene mayores resistencias periféricas, menor volumen plasmático, menor frecuencia y contractilidad cardiacas, menor gasto cardiaco, disminución de la sensibilidad de los basorreceptores, menor producción de prostaglandinas renales y mayores niveles de catecolaminas circulantes. A todo ello hay que añadir la disminución de la distensibilidad de las grandes arterias y el aumento de la masa ventricular, que conduce a una mayor prevalencia de hipertrofia ventricular izquierda, así como al progresivo deterioro de la función renal por nefroangiosclerosis.

4.2.4.2.2. Hipertensión y Diabetes.

También la diabetes se encuentra presente en los diferentes modelos estimados; es conocida la asociación de hipertensión y diabetes. Hasta un 17,5% de los sujetos de entre 65 y 80 años y el 15% de los mayores de 80 años hipertensos, son diabéticos. Este hecho es mas frecuente en las mujeres que en los hombres.

El desarrollo de la HTA de los sujetos diabéticos, no tiene un mecanismo patogénico claro y probablemente se deba al efecto conjunto de factores como: la resistencia insulínica, la función renal, la homeostasis del sodio, la pared vascular

o el control de la presión arterial en la diabetes. En cualquier caso la HTA aparece con una mayor frecuencia en los pacientes diabéticos y se calcula que su prevalencia es entre 1,5 y dos veces superior a la de los no diabéticos. La hipertensión y la diabetes se consideran dos factores importantes de riesgo de enfermedad vascular. La asociación de estos dos factores en un mismo individuo incrementa las posibilidades de cardiopatía isquémica, enfermedad cerebrovascular, enfermedad vascular periférica, entre otras patologías, al disminuir el umbral de riesgo.

Si bien el efecto de la diabetes no es tan potente en la génesis de la enfermedad cerebrovascular aguda como en la cardiopatía isquémica o en la arteriopatía periférica, la asociación existe pudiendo el tratamiento intensivo de la diabetes disminuir el riesgo de padecer un ictus. En este sentido resulta especialmente eficaz, como estrategia de prevención primaria del ictus en los sujetos con diabetes el tratamiento de la HTA.

La afectación de los miembros interiores derivada de la insuficiencia vascular periférica secundaria a la diabetes, tiene un origen multifactorial. Cuando aparece suele haber enfermedad vascular severa y suelen estar afectados ya la circulación coronaria, renal y cerebrovascular. La mitad de las amputaciones de miembros inferiores se realiza en diabéticos, en los que el riesgo de sufrirlas esta aumentado entre 10 y 24 veces respecto a la población general. En la patogenia, están implicados múltiples factores como: daño vascular, afectación del sistema nervioso periférico e infección. La neuropatía periférica disminuye la percepción del dolor, con la consiguiente facilidad para sufrir heridas en la piel y atrofia muscular, con lo que cambian los puntos de apoyo del pie, descargando la presión sobre zonas no preparadas para ello, lo que facilita la aparición de úlceras y de infecciones.

4.2.4.2. Corrección de la sobreestimación del vector beta de coeficientes de regresión logística.

Cuando los datos son escasos, vulnerando claramente la regla uno diez, no entran en el modelo todas las variables que realmente contiene, aún más, si considerásemos variables independientes, con un número suficiente de datos todas entrarían en el modelo. Esto lo demuestra la simulación realizada, epígrafe 2.9.6.5. Sin embargo, esto no quiere decir que exista una correlación positiva entre el número de variables en el modelo y el volumen de datos; pues con los datos suficientes entrarían en el modelo las variables que realmente forman parte del mismo, y al aumentar el volumen de datos no entran nuevas variables.

En parecidos términos nos podríamos expresar sobre la inflación del vector beta. A medida que el volumen de datos va disminuyendo, siempre violando la regla 1-10, mayor inflación y error presentan las componentes del vector beta. A un volumen de datos tal que ya estén contenidas en el modelo todas las variables que realmente deben de entrar; aumentando el número de datos, la variación producida en el valor de los coeficientes y su error es mínima.

Respecto a los Factores de Corrección, los resultados obtenidos son perfectamente consecuentes con los valores del vector beta. Cuantos menos datos se usen en la generación del modelo, mayor es la inflación de beta y más decrecen los factor de corrección. Los citados factores manifiestan su utilidad, en situaciones de una importante escasez de datos. Realizan correcciones muy drásticas, situando el vector transformado más cercano al estimado sin escasez de datos, y por tanto proporciona unas odd-ratios más aproximadas, eliminando su inflación, y obteniendo nuevos coeficientes a la baja. No es necesaria su aplicación cuando la reducción de datos no es importante. De los dos factores, el corregido proporciona una mejor aproximación que el puro. Hemos de concluir que aplicando los factores de corrección se obtienen un vector beta y unas odd-ratios mucho mas aproximadas a los valores reales que si la corrección no se hubiese efectuado, lo que pone de relieve la bondad de esta metodología.

Apéndice I. Glosario.

Se explican en este glosario términos/expresiones empleadas en el capítulo II (Métodos Estadísticos), en referencia a infinitésimos usados en desarrollos matemáticos realizados en el mencionado capítulo.

* Dadas dos sucesiones $\{x_n\}_{n\in\square}$ e $\{y_n\}_{n\in\square}$, decimos que y_n *es un infinitésimo de orden o de* x_n, esto es $y_n = o(x_n)$, si la sucesión cociente converge a cero. En términos más formales: $\forall \varepsilon > 0$, y $\exists n_0 \in \square$, tal que $\forall n > n_0$ se tienen que: $\left|\frac{y_n}{x_n}\right| \leq \varepsilon$.

* Dadas dos sucesiones $\{x_n\}_{n\in\square}$ e $\{y_n\}_{n\in\square}$, decimos que y_n *es un infinitésimo de orden O de* x_n, esto es $y_n = O(x_n)$, si la sucesión cociente está acotada. En términos más formales: $\exists C > 0$, y $\exists n_0 \in \square$, tal que $\forall n > n_0$ se tienen que: $\left|\frac{y_n}{x_n}\right| \leq C$.

* Sea una sucesión numérica $\{x_n\}_{n\in\square}$, e y_n una variable aleatoria, decimos que y_n *es de orden* O_p *de* x_n, esto es $y_n = O_P(x_n)$, si $\forall \eta > 0$, $\exists (C > 0) \in \square$ y $\exists n_0 \in \square$, tal que $\forall n > n_0$ se tiene que: $\mathbf{P}\left(\left|\frac{y_n}{x_n}\right| \leq C\right) \geq 1 - \eta$.

Apéndice II. Programas informáticos.

I. Introducción.

En este apéndice se muestran los códigos fuentes de los programa informáticos que implementan algunas de las metodologías expuestas en esta memoria. La primera herramienta elaborada contiene la implementación de las transformaciones de ByC , conjuntamente para casos y controles; se ha realizado en Pascal 7.0. La segunda es la implementación del bootstrap, como método de corrección para RL; se ha implementado en C++ Builder 5 y usa bibliotecas NAG Mark6. También se ha realizado, en el entorno, C++ Builder 5 la implementación de curvas ROC para diversa metodologías. Por último y en el mismo entorno anterior se implementa un herramienta comparativa de dos medias, usando el t-test y el contraste de las permutaciones bootstrap.

Los resultados obtenidos al procesar la base de datos con estos programas, se encuentran expuestos en el capítulo III.

Para mayor claridad, en los listados se han conservado mensajes al usuario no imprescindibles y puntos de control, que figuran como comentarios.

En el CD-ROM que acompaña a esta memoria se encuentran los códigos fuente y objeto de los programas reseñados, así como las bases de datos utilizadas.

I.1. Comparación de dos medias.

El programa implementa la comparación de dos medias, procedentes de muestras independientes. Utiliza dos metodologías, la primera de ellas es el t-test, donde se estima la cantidad pivotal *T*. El segundo es el test de las permutaciones bootstrap, que no supone normalidad, ni aplica el teorema central del límite, se basa

en la autosuficiencia de las réplicas bootstrap. El entorno utilizado es el C++ Builder 5.

La complejidad del algoritmo es del orden O((*B*+1)**n*), con *B*= número de réplicas bootstrap y *n* longitud del fichero.

```
/*****************************************************************
*
/*                                                    *
/*          COMPARACIÓN DE MEDIAS                       *
/*        (CORRECCIÓN PIVOTAL Y BOOTSTRAP)              *
/*                                                    *
/*****************************************************************
**/

#include <stdlib.h>      //Nºs aleatorios
#include <stdio.h>
#include <math.h>        //Funciones matemáticas
#include <iostream.h>
#include <fstream.h>     //Flujo de archivos
#include <conio.h>       //Coordenadas de pantalla
#include <string.h>
#include <time.h>

const int nmax=3000;
const int nmax2=30000;
const float cc=10000;
const float pi=3.141528;

double gauss(double x)
 {
  double f;
  f=(1/sqrt(2*pi))*exp(-x*x/2);
  return f;
 }

double pvalor(double t)
 {
  int i,R;
  float d,a,s;
  R=10000;
  d=0.001;
  a=t;s=0;
  for(i=1;i<=R;i++)
   {
    a=a+d;
```

```
  s=s+d*(gauss(a)+gauss(a-d))/2;
 }
s=2*s;
return s;
}

double ranuni(void)
{
 double u,z;
 u=random(cc);
 z=(u+1)/cc;
 return z;
}

float normal(float mu, float sigma)  //Genera una v.a. N(mu,sigma)
{
 double u,v,z;
 u=random(cc);u=(u+1)/cc;
 v=random(cc);v=(v+1)/cc;
 z=sqrt(-2*log(u))*cos(2*pi*v);
 z=mu+sigma*z;
 return z;
}

float expo(float a)
{
 float v,z;
 v=ranuni();
 z=-a*log(v);
 return z;
}

float media(int n,float *y)
{
 int i;
 float z;
 z=0;
 for(i=1;i<=n;i++)
 z=z+y[i];
 z=z/float(n);
 return z;
}

float varianza(int n,float *y)
{
 int i;
 float mu,z;
 mu=media(n,y);
```

```
 z=0;
 for(i=1;i<=n;i++)
 z=z+pow(y[i]-mu,2)/float(n-1);
 return z;
 }


float skew(int n,float *y)

 {
 int i;
 float z;
 z=0;
 for(i=1;i<=n;i++)
 z=z+pow(y[i]-media(n,y),3);
 z=z/float(n);
 return z;
 }

double spoole(int n1,double v1,int n2,double v2)
 {
 float s;
 s=((n1-1)*v1+(n2-1)*v2)/double(n1+n2-2);
 s=sqrt(s)*sqrt(1/double(n1)+1/double(n2));
 return s;
 }


void ordena(long int n,float *x)
 {
 register int i,j;
 float a,b;
 for(i=1;i<=n-1;i++)
 for(j=i+1;j<=n;j++)
  {
   a=x[i];
   b=x[j];
   if(a>b)
    {
     x[i]=b;
     x[j]=a;
    }
  }
 }


void main()      //programa principal
 {
```

```
 register int i,l;
 int op,od,n,n1,n2,u,con;
 long int B,lu1,lu2;
 float mu,sigma;
 float
g1,g2,m1,m2,dm,sd1,sd2,*g,*y,*y1,*y2,*yb,mb1,mb2,dmb,pval,*Z,v1,v2,vb1,vb
2;
 double se,p,seb,t,low,upp;
 char fi[40];
 FILE *f;

 f=fopen("tstu.TXT","w");
 g=NULL;
 g=new float[nmax];
 y=NULL;
 y=new float[nmax];
 y1=NULL;
 y1=new float[nmax];
 y2=NULL;
 y2=new float[nmax];
 yb=NULL;
 yb=new float[nmax];
 Z=NULL;
 Z=new float[nmax2];

 gotoxy(15,2);
 cout<<"****** COMPARACION DE DOS MEDIAS ******";
 gotoxy(3,4);
 printf("Datos:  1=LECTURA DE ARCHIVO EXTERNO ; 2=SIMULADOS =>
");
 scanf("%d",&op);
 fflush(stdin);

 if(op==1)    //Lectura de archivo externo
  {
  gotoxy(3,5);
  printf("Fichero de datos ==> ");
  gets(fi);
  f=fopen(fi,"r");
  if (!f) printf("No abrió el fichero");

  n=0;
  while(!feof(f))
   {
    n++;
    gotoxy(60,4);
    printf(" --> N = %d",n);
```

```
   fscanf(f,"%f %f\n",&g[n],&y[n]);
  }
 fclose(f);
}
else //Simulación de datos
{
 randomize();
 gotoxy(3,5);
 printf("Grupo A (1=Normal;2=Exp.) = ");
 scanf("%d",&od);
 gotoxy(33,5);
 printf(" -> N = ");
 scanf("%d",&n1);
 gotoxy(48,5);
 printf("Media = ");
 scanf("%f",&mu);
 if(od==1)
  {
   gotoxy(60,5);
   printf("SD = ");
   scanf("%f",&sigma);
  }
 for(i=1;i<=n1;i++)
  {
   g[i]=1;
   if(od==1) y[i]=normal(mu,sigma);else y[i]=expo(mu);
  }
 gotoxy(3,6);
 printf("Grupo B (1=Normal;2=Exp.) = ");
 scanf("%d",&od);
 gotoxy(33,6);
 printf(" -> N = ");
 scanf("%d",&n2);
 gotoxy(48,6);
 printf("Media = ");
 scanf("%f",&mu);
 if(od==1)
  {
   gotoxy(60,6);
   printf("SD = ");
   scanf("%f",&sigma);
  }
 for(i=n1+1;i<=n1+n2;i++)
  {
   g[i]=2;
   if(od==1) y[i]=normal(mu,sigma);else y[i]=expo(mu);
  }
 n=n1+n2;
```

```
 }

g1=g[1];
i=1;
do
 {
  i=i+1;
  g2=g[i];
 }
while(g2==g1 && i<=n);

n1=0;n2=0;
for(i=1;i<=n;i++)
 {
  if(g[i]==g1)
   {
    n1=n1+1;
    y1[n1]=y[i];
   }
  else
   {
    n2=n2+1;
    y2[n2]=y[i];
   }
 }

m1=media(n1,y1);
m2=media(n2,y2);
dm=fabs(m1-m2);
v1=varianza(n1,y1);sd1=sqrt(v1);
v2=varianza(n2,y2);sd2=sqrt(v2);
se=spoole(n1,v1,n2,v2);
t=(m1-m2)/se;
p=pvalor(fabs(t));
gotoxy(3,8);
printf("Grupo  N    Media       Desv. Est.");
gotoxy(3,10);
printf("%3.0f %4d  %10.5f   %10.5f",g1,n1,m1,sd1);
gotoxy(3,11);
printf("%3.0f %4d  %10.5f   %10.5f",g2,n2,m2,sd2);
gotoxy(3,13);
printf(" Prueba de la t de Student:  T = %8.3f   g.l.=%d   P=%8.5f",t,n1+n2-
2,p);

//Test de permutaciones Bootstrap
randomize();
do
 {
```

```
  gotoxy(3,15);
  printf("Num. de replicas bootstrap (Recomiendo %d) = ",10*n);
  scanf("%d",&B);
 }
while(B>=nmax2);
con=0;
for(l=1;l<=B;l++)
 {
  for(i=1;i<=n;i++)
   {
    u=random(n)+1;
    yb[i]=y[u];
   }
  mb1=0;
  for(i=1;i<=n1;i++)
  mb1=mb1+yb[i]/n1;
  mb2=0;
  for(i=n1+1;i<=n;i++)
  mb2=mb2+yb[i]/n2;
  dmb=fabs(mb1-mb2);
  fprintf(f,"%8.3f\n",dmb);
  if(dmb>dm) con=con+1;
 }
pval=(1+con)/(1+float(B));
fclose(f);

gotoxy(3,16);
printf("Test bootstrap de permutaciones ; p-valor = %8.5f",pval);

//Intervalos de confianza bootstrap
for(l=1;l<=B;l++)
 {
 for(i=1;i<=n1;i++)
  {
   u=random(n1)+1;
   yb[i]=y1[u];
  }
 mb1=media(n1,yb);
 vb1=varianza(n1,yb);
 for(i=1;i<=n2;i++)
  {
   u=random(n2)+1;
   yb[i]=y2[u];
  }
 mb2=media(n2,yb);
 vb2=varianza(n2,yb);
 seb=spoole(n1,v1,n2,v2);
 Z[l]=(mb1-m1-mb2+m2)/seb;
```

```
 fprintf(f,"%8.3f   %8.3f  %8.3f\n",mb1,mb2,seb);
 }

f=fopen("tboot.TXT","w");
ordena(B,Z);
for(l=1;l<=B;l++)
fprintf(f,"%8.3f\n",Z[l]);
fclose(f);

f=fopen("datos.dat","w");
for(i=1;i<=n;i++)
fprintf(f,"%6.0f  %8.2f\n",g[i],y[i]);
fclose(f);

lu1=int(0.025*B);
lu2=int(0.975*B);
low=dm+Z[lu1]*se;
upp=dm+Z[lu2]*se;
gotoxy(3,17);
printf("Diferencia de medias = %f  SE = %f  ",-dm,se);
gotoxy(3,18);
printf("Diferencia de medias: 95IC bootstrap ==> %8.3f;%8.3f",-upp,-low);
gotoxy(50,23);
printf("P.Saavedra");
gotoxy(38,24);
system("pause");
}
```

I.2. Transformaciones de Box y Cox.

Esta nueva herramienta está realizada en lenguaje PASCAL (Borland Pascal 7.0), si bien el algoritmo es fácilmente exportable a cualquier otro entorno. Las transformadas inversas se han calculado con otra herramienta elaborada para la ocasión; los códigos fuente se encuentra en este epígrafe. El programa realiza, bajo hipótesis de homoscedasticidad, las transformaciones de Box y Cox, para buscar normalidad simultánea en dos conjuntos de datos: casos y controles. Necesita ficheros de entrada de casos y controles, y produce otros dos: resulc.dat y resultc.dat (ASCII), con los datos ya normalizados, que son la transformación óptima, la cual maximiza la función de log verosimilitud. También entrega el valor óptimo de

lambda, potencia de BOX y COX. Por último calcula un intervalo de confianza para el valor del parámetro lambda.

COMPLEJIDAD: Si llamamos *n* a la suma de las longitudes del fichero de casos, más el de controles; la complejidad del algoritmo es aproximadamente del orden de 82**n*; esto es, proporcional al valor de *n*; lineal ==→ O(*n*).

Código fuente:

```
Program BoxCox (input,output);
uses crt,printer,dos;
type valores=array[1..2,1..81] of real;

var casos,controles,resulc,result:text; lambdav:valores;
   c:1..82; frc,frt:text;
   mmc,mmct:real;
   varc,varct:real;
 l1,l2,n1,n2,n3:integer;
   ca:char;
   b,bb:boolean;
  s1,s2:string;
(*Realiza una corrección por continuidad *)

Procedure CPC (var ffcc,fftt,auxff1,auxff2:text);
  var rr1:real;

begin
 reset(ffcc);
 reset(fftt);
 rewrite(auxff2);
 rewrite(auxff1);
 while not eof(ffcc) do begin
                read(ffcc,rr1);
                rr1:=rr1+0.6;
                rr1:=int(round(rr1));
                write(auxff1,rr1)
                end;

 while not eof(fftt) do begin
                read(fftt,rr1);
                rr1:=rr1+0.6;
                rr1:=int(round(rr1));
                write(auxff2,rr1)
```

```
                end;
rewrite(ffcc);
rewrite(fftt);
reset(auxff1);
reset(auxff2);

while not eof(auxff1) do begin
                read(auxff1,rr1);
                write(ffcc,rr1)
                end;

while not eof(auxff2) do begin
  read(auxff2,rr1);
  write(fftt,rr1)
  end;
reset(ffcc);
reset(fftt);
rewrite(auxff2);
 rewrite(auxff1)
  end;

  (*Calcula la transformada de cada dato, para una potencia no nula
 *)

Procedure Transformar  (lambda1:real; var cc,cc1:text;var ll1:integer);
  var cont1:integer; rr1:real;

  begin
   reset(cc);
   rewrite(cc1);
   cont1:=0;

   while not eof(cc) do begin
                cont1:=cont1+1;
                read(cc,rr1);
                rr1:=Ln(rr1);
                rr1:=lambda1*rr1;
                rr1:=exp(rr1);
                rr1:=((rr1)-1)/lambda1;
                write(cc1,rr1)
                 end;

    ll1:=cont1
  end;

  (* Calcula la transformada para el valor cero de la potencia*)

  Procedure Transcero  (var cc,cc1:text);
```

```
var rr1:real;

begin
     reset(cc);
     rewrite(cc1);
     while not eof(cc) do begin
                    read(cc,rr1);
                    rr1:=ln(rr1);
                    write(cc1,rr1)
                              end

 end;

Function Media (var cc:text;ll1:integer):real;
var rr1,vv:real;
begin
  reset(cc);
  vv:=0;
  while not eof(cc) do begin
                    read(cc,rr1);
                    vv:=vv + rr1
                    end;
  Media:=vv/ll1
  end;

Function Semivar (med1:real; var cc:text):real;
var rr1,vv:real;

begin
 reset(cc);
 vv:=0;
 while not eof(cc) do begin
                   read (cc,rr1);
   vv:=vv+sqr(rr1-med1)
   end;
   Semivar:=vv
   end;

  Function Log (var cc:text):real;
  var rr1,vv:real;

  begin
   reset(cc);
   vv:=0;

   while not eof(cc) do begin
                    read (cc,rr1);
                    vv:=vv+Ln(rr1)
```

```
                end;
  Log:=vv
  end;

 (*Calcula, para una transformación dada, la función de
    verosimilitud*)

 Procedure logvero (var casos1,resulc1,controles1,resulct1:text;
var
lambdav1:valores;ll1,ll2:integer;var b:boolean);
 var mmc1,mmct1,varc1,varct1,min,logvero1:real;
     nn1,nn2:integer;

 begin
       mmc1:=Media(resulc1,ll1);
       mmct1:=Media(resulct1,ll2);
       varc1:=Semivar(mmc1,resulc1)/ll1;
       varct1:=Semivar(mmct1,resulct1)/ll2;
     if sqrt(varc1)<sqrt(varct1) then min:=sqrt(varc1)
                        else min:=sqrt(varct1);
     if int(varct1)=int(varc1) then begin   (*Solo se consideran
     soluciones HOMOSCEDASTICAS*)
     b:=false;
     logvero1:=-((ll1+ll2)/2)*Ln(2*pi*varc1);
     logvero1:=logvero1+((lambdav1[1,c]-
     1)*(Log(casos1)+Log(controles1)));
     logvero1:=logvero1-(1/(2*varc1))*((ll1*varc1)+(ll2*varct1));
     lambdav1[2,c]:=logvero1                end

     else lambdav[2,c]:=-1000000000;
     gotoxy (10,10);

     if lambdav[2,c]=-1000000000 then write(lambdav[1,c],' ','
     Sin valor (No presenta Homoscedasticidad)')
     else write(lambdav[1,c],'       ', lambdav[2,c]);
     end;


(* Calcula el intervalo de confianza para lambda, potencia de BOX y COX*)

 Procedure Intconf (vv1:valores; nc1,nn1:integer; var
mm1,mm2:real;var bb1,bb2:boolean);
 var  c1:integer;rr,rr1:real;
  begin
    case nc1 of
       95: rr:=3.841;
       98: rr:=5.412;
       99: rr:=6.635;
```

```
 end;
 bb1:=true;
 bb2:=true;
 rr:=-(rr/2);
 rr:=rr+vv1[2,nn1];
 c1:=nn1;
 repeat
if ((c1>0) and  (vv1[2,c1]<>-1000000000)) then  begin
          if  vv1[2,c1]=rr then begin
                        mm1:=vv1[1,c1];
                        bb1:=not bb1
                          end;
          if ((vv1[2,c1]>rr) and (vv1[2,c1-1]<rr)) then
   if ((vv1[2,c1]<>-1000000000) and (vv1[2,c1-1]<>
    -1000000000)) then
             begin
             mm1:=vv1[1,c1-1]-vv1[1,c1];
             rr1:=vv1[2,c1-1]-vv1[2,c1];
             mm1:=mm1/rr1;
             mm1:= mm1*(rr-vv1[2,c1]);
             mm1:= mm1+vv1[1,c1];
             bb1:=not bb1
              end;

                    end;
 c1:=c1-1;

 until ((c1=1) or (not bb1));
 if vv1[2,1]=rr then  mm1:=vv1[1,1];
 c1:=nn1;
 repeat
  if ((c1<81) and  (vv1[2,c1]<>-1000000000)) then  begin
            if  vv1[2,c1]=rr then begin
                        mm2:=vv1[1,c1];
                        bb2:=not (bb2)
                          end;
            if ((vv1[2,c1]>rr) and (vv1[2,c1+1]<rr)) then
                        if ((vv1[2,c1]<>-1000000000) and
                        vv1[2,c1+1]<>-1000000000)) then
             begin
             mm2:=vv1[1,c1+1]-vv1[1,c1];
             rr1:=vv1[2,c1+1]-vv1[2,c1];
             mm2:=mm2/rr1;
             mm2:= mm2*(rr-vv1[2,c1]);
             mm2:= mm2+vv1[1,c1];
             bb2:=not (bb2)
              end;
```

```
                    end;
    c1:=c1+1;

    until ((c1=81) or (not bb2));
    if vv1[2,81]=rr then  mm2:=vv1[1,81]
     end;


 Begin (*Comienza el programa principal*)

textbackground(6);
clrscr;
     gotoxy(5,6);
     highvideo;
     write(' TRANSFORMACIONES DE BOX Y COX, PARA
           ESTUDIOS DE CASOS Y CONTROLES.');
     lowvideo;
     gotoxy(5,8);
     write('     Este programa realiza, bajo hipótesis de
              homoscedasticidad, ');
     gotoxy(5,10);
     write(' las transformaciones de BOX y COX, par buscar
     normalidad simultanea');
     gotoxy(5,12);
     write(' en dos conjuntos de datos: casos y controles.');
     gotoxy(5,14);
     write('     Necesita  ficheros de entrada de casos y controles');
     gotoxy(5,16);
     write(' y produce otros dos: resulc.dat y resultc.dat (ASCII),con los
      datos');
     gotoxy(5,18);
     write(' ya normalizados, que son la transformación óptima, la cual
     maximiza ');
     gotoxy(5,20);
     write(' la función de logverosimilitud. También da el valor óptimo
     de lambda,');
     gotoxy(5,22);
     write (' potencia de BOX y COX.');
     gotoxy(5,24);
     write('     Por último calcula un intervalo de confianza para el
              valor ');
     gotoxy(5,25);
     write('  del par metro lambda.');
     writeln;
     write('          Pulse una tecla para ejecutar el programa.....');

     repeat until keypressed;
     ca:=readkey;
```

```
clrscr;
gotoxy(5,6);
highvideo;
write(' TRANSFORMACIONES DE BOX Y COX, PARA
      ESTUDIOS DE CASOS Y CONTROLES.');
lowvideo;

gotoxy(2,10);

write('Introduzca nombre  del fichero de casos ======>');
readln(s1);
assign(casos,s1);
gotoxy(2,12);
write('Introduzca nombre del fichero de controles =====>');
readln(s2);
assign(controles,s2);
reset(casos);
reset(controles);
assign(resulc,'C:\datos\frc.dat');
rewrite(resulc);
assign(result,'C:\datos\frct.dat');
rewrite(result);
clrscr;
gotoxy(5,6);
highvideo;
write(' TRANSFORMACIONES DE BOX Y COX, PARA
      ESTUDIOS DE CASOS Y CONTROLES.');
lowvideo;
b:=true;
repeat
while not ((eof(casos)) or (b=not b)) do begin
                        read(casos,mmc);
                        if mmc=0 then b:=not b
                        end;

 while not ((eof(controles) or (b=not b))) do begin
                        read(controles,mmc);
                        if mmc=0 then b:=not b
                        end;
 until ((b=not b) or ((eof(casos) and (eof(controles)))));
if b=false then begin
CPC(casos,controles,resulc,result);
writeln;
write('     Se ha realizado correcci¢n por continuidad')
end;
 lambdav[1,1]:=-4;
 for c:=2 to 31 do lambdav[1,c]:=lambdav[1,c-1]+0.1;
 lambdav[1,32]:=-0.9;
```

```
for c:=33 to 39 do lambdav[1,c]:=lambdav[1,c-1]+0.1;
lambdav[1,40]:=-0.1;
lambdav[1,41]:=0;
for c:=42 to 81 do lambdav[1,c]:=lambdav[1,c-1]+0.1;
b:=true;
gotoxy(10,8);
write('Valor de lambda.',' ',' Valor de la función
                                Logverosimilitud.');

for c:=1 to 81 do if lambdav[1,c]=0 then begin
          Transcero(casos,resulc);
         Transcero(controles,result);
         Logvero(casos,resulc,controles,result,lambdav,l1,l2,b)
                  end
              else begin
  Transformar(lambdav[1,c],casos,resulc,l1);
  Transformar(lambdav[1,c],controles,result,l2);
  Logvero(casos,resulc,controles,result,lambdav,l1,l2,b)
                  end;

(* Busca la transformación que maximiza la función
  log verosimilitud*)
n1:=1;
for c:=1 to 80 do if lambdav[2,c]<lambdav[2,c+1] then n1:=c+1;
if lambdav[2,n1]<lambdav[2,81] then n1:=81;
if lambdav[1,n1]=0 then begin
                  Transcero(casos,resulc);
                  Transcero(controles,result)
                  end
               else begin
Transformar(lambdav[1,n1],casos,resulc,l1);
Transformar(lambdav[1,n1],controles,result,l2)
                  end;
gotoxy(4,12);
highvideo;
sound(500);
nosound;
if b=true then writeln('  NINGUNA SOLUCION PRESENTA
                          HOMOSCEDASTICIDAD')
   else begin
    writeln('     LA SOLUCIÓN ÓPTIMA DE MÁXIMO ES
                  LAMBDA=', ' ',lambdav[1,n1]:6);
    writeln('       EL VALOR MÁXIMO DE LA FUNCIÓN ES=
                  ',lambdav[2,n1])
   end;
writeln;
lowvideo;
writeln('       N° DE CASOS PROCESADOS: ',l1,'  N° DE
```

```
                CONTROLES PROCESADOS: ',l2);
 writeln;
 writeln;

 writeln('     " Quiere imprimir estos resultados ?. (S/N)...');
 gotoxy(53,18);
 ca:=readkey;

 if upcase(ca)='S' then begin
    for c:=1 to 81 do if lambdav[2,c]=-1000000000 then
 Writeln(lst,lambdav[1,c]:6,'  No presenta Homoscedasticidad.')
 else writeln(lst,lambdav[1,c]:6,'        ', lambdav[2,c]);


    if b=true then writeln(lst,'  NINGUNA SOLUCION
                    PRESENTA  HOMOSCEDASTICIDAD')
    else begin

     writeln(lst,' LA SOLUCION OPTIMA DE MAXIMO ES
                 LAMBDA=', ' ',lambdav[1,n1]:6);

     writeln(lst,'   EL VALOR MAXIMO DE LA FUNCION ES=
                  ',lambdav[2,n1])

   end;

     writeln(lst,'    N° DE CASOS PROCESADOS: ',l1,'  N° DE
                    CONTROLES PROCESADOS: ',l2);
               end;
     writeln;
clrscr;
gotoxy(5,6);
highvideo;
write(' TRANSFORMACIONES DE BOX Y COX, PARA
      ESTUDIOS DE CASOS Y CONTROLES.');
lowvideo;
gotoxy(4,10);
write('" Desea hallar un intervalo de confianza para el parámetro
       lambda? (S/N)');
ca:=readkey;
if upcase(ca)='S'  then begin

        repeat
        clrscr;
        gotoxy(5,6);
        highvideo;
write(' TRANSFORMACIONES DE BOX Y COX, PARA
        ESTUDIOS DE CASOS Y CONTROLES.');
```

```
                    lowvideo;
                    gotoxy(6,12);

write('"Debe elegir el nivel de confianza. Este debe ser 95%,98% o
      99% ==>');
                     gotoxy(73,12);
                     read(n2);
        readln;
        until ((n2=95) or (n2=98) or (n2=99));

        Intconf(lambdav,n2,n1,mmc,mmct,b,bb);
        writeln;
        writeln;
if ((b) and (bb)) then  write('       IMPOSIBLE  CALCULAR
                                    INTERVALO DE CONFIANZA');
if ((b) and (not bb)) then  begin
                  writeln('     IMPOSIBLE  CALCULAR
                              SUBINTERVALO A LA IZQUIERDA.');
                  writeln;
                  writeln('"Subintervalo de confianza, para el
                        parámetro lambda= ',lambdav[1,n1]:6);
                  writeln(' al  ',n2 ,'  %  es:      (   ,',mmct:9,' )' );
                  writeln
                     end;

if ((not b) and (bb)) then  begin
                  writeln('     IMPOSIBLE  CALCULAR
                              SUBINTERVALO A LA DERECHA.');
                  writeln;
                  write('"Subintervalo de confianza, para el
                 parámetro lambda= ',lambdav[1,n1]:6,' al ',n2 ,' %');
                  writeln;
                  writeln('         es:        ( ',mmc:9,',    )');
                  writeln
                     end;

if ((not b) and (not bb)) then  begin
                  write('"El intervalo de confianza, para el par metro
                        lambda',lambdav[1,n1]:6,' al ',n2 ,'% es:');
                  writeln;
                  writeln('           ( ',mmc:9,' , ',mmct:9,')');
                  writeln
                     end;

            end;

Close(casos);
Close(controles);
```

```
  close(resulc);
  close(result);

  (*Final del programa*)
  gotoxy(40,24);
  write('Pulse cualquier tecla para terminar ');
  ca:=readkey;
  clrscr;
  gotoxy(20,20);
  highvideo;
  write('Programa de Bioestadística. Autor E. Núñez ');
  delay(2500);
  lowvideo;
  clrscr
End.
```

Código fuente del programa de transformación inversa puntual:

Este programa realiza el cálculo de las inversas de las transformaciones de BOX y COX, usadas para buscar normalidad. Pide, como entrada, el valor cuyo inverso se va a calcular y la potencia, lambda, de ByC. Entregando como salida el original del valor. La función inversa de la familia existe por ser monótona creciente.

```
Program InvByC (input,output);
uses crt;
var r1,r2,aux:extended; ca:char;
procedure Invertir (var rr1,rr2:extended);

begin
 if rr2<>0 then begin
                rr1:=rr1*r2+1;
                rr1:=exp((1/rr2)*Ln(rr1))
                end

             else rr1:=exp(rr1);
 end;
begin (*Programa principal*)
  clrscr;
  textbackground(6);
    gotoxy(5,6);
    highvideo;
    write('     INVERSA DE LAS TRANSFORMACIONES DE BOX Y
                COX.');
```

```
lowvideo;
gotoxy(5,8);
write(' Este programa realiza el cá lculo de las inversas de ');
gotoxy(5,10);
write(' las transformaciones de BOX y COX,usadas para buscar
      normalidad. ');
gotoxy(5,12);
write('   Pide, como entrada el valor cuyo inverso se  va a
        calcular.   ');
gotoxy(5,14);
write('y la potencia,lambda, de BOX y COX. Entregando como
      salida el original ');
gotoxy(5,16);
write('del valor. La función inversa de la familia de Box y Cox,
      existe siempre');
gotoxy(5,18);
write('por ser monótona creciente.  ');
writeln;
writeln;
writeln;
write('          Pulse una tecla para ejecutar el programa.....');
repeat until keypressed;
ca:=readkey;
repeat
clrscr;
gotoxy(5,6);
highvideo;
write('      INVERSA DE LAS TRANSFORMACIONES DE BOX Y
            COX.');
lowvideo;
gotoxy(8,8);
write(' Escriba el valor cuya inversa quiere calcular ===>');
gotoxy(8,60);
read(r1);
gotoxy(8,10);
writeln(' Escriba el valor de lambda para la transformaci¢n ===>');
gotoxy(63,10);
read(r2);
Invertir (r1,r2);
gotoxy(14,14);
highvideo;
write(' El valor transformado es:  ',r1);
lowvideo;
gotoxy(16,16);
write(''' Desea realizar más calculos ?. (S/N)===>');
ca:=readkey
until not ('S'=upcase(ca));
lowvideo;
```

```
 gotoxy(40,24);
 write('Pulse cualquier tecla para terminar ');
ca:=readkey;
clrscr;
gotoxy(20,20);
highvideo;
write('Programa de Bioestadística. Autor E. Nunnez ');
delay(2500);
lowvideo;
clrscr
end.
```

I.3. Curvas ROC.

Este programa implementa el análisis de bondad de un protocolo o proceso diagnóstico, basándose en estimaciones de curvas ROC. Considera tres variantes metodológicas: estimación cruda, otra basada en el supuesto de normalidad y por último, estimación no paramétrica. En cada caso se calculan diversos puntos cut-off, intervalo y área bajo las curvas. Es necesario un fichero de entrada con resultados diagnósticos; el programa produce otros tres, que albergan los datos (pares de falsos positivos y sensibilidad) para la las tres diferentes metodologías.

El algoritmo de cálculo del Bandwidth óptimo, es doble, pues además de la Regla del Pulgar (2.6.12), se implementa un método de validación cruzada (máxima verosimilitud) , que estima muy exactamente el ancho de la ventana; esta precisión se refleja en un aumento considerable del tiempo de ejecución, que es cinco veces n al cubo.

La complejidad del algoritmo es, ejecutando los tres métodos, y designado por n la longitud del fichero de entrada, es del orden ======➔ $o(3n + 5n^3)$

Ficheros de salida:

Estimación real:

Datos_ROC_real.TXT

Estimación normal:

Datos_ROC_nor.TXT

Estimación no paramétrica:

Datos_ROC_nopar.TXT

Código fuente:

```
//---------------------------------------------------------------------------
 /* Programa de estimación de curvas ROC.

            Autor: E. Núñez. */


#include <vcl.h>
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#include <iostream.h>
#include <fstream.h> //flujo de archivos
#include <conio.h>  //coordenadas de pantalla
#include <memory.h> // matrices dinámicas.
#include <time.h>
#include <dos.h>

 const long ll1=2000;
 const long ll2=4;
 const long ll3=1000;
 const double maxint=9e+15;

  void tabla_normal (double tt_nn[35][10],FILE *f_normal);

  //A continuación viene el nucleo de Epanechnikov.

  double kernel_Epa (double hh,double dd_oo[ll1][ll2],double vv,double cctt)
      { int f1,c1;
        int cont11=1;
        double acum1=0,acum2=0;
        c1=-1;
        f1=0;
        do
```

```
        {
        if (vv!=dd_oo[f1][1]) f1++;
        else c1=f1;
         }
        while ((f1<ll1) && (c1!=-1));


         for (f1=c1;f1<ll1;f1++)
         {if ((vv!=dd_oo[f1][1])&& (cctt==dd_oo[f1][0]))
            { cont11++;
             acum1=0;
             acum1=(vv- dd_oo[f1][1])/hh;
             if (fabs(acum1)<=1)
              { acum1=0.75*(1-(pow(acum1,2)));
                acum2+=acum1;
              }
            }
          }
        if (cont11!=0)return acum2/double(cont11);
         else return 0;
       }

double c_kernel_Epa (double xx1)
  {double acum1=0;

   if (xx1<-1) acum1=0;
   if (fabs(xx1)<=1) acum1=(2+3*xx1-pow(xx1,3))/4.0;
   if (xx1>1) acum1=1;
   return acum1;

  }


 // Nucleo integrado de Epanechnikov.

 void k_integrado_Epa (double ROC1[ll3][3],double hh,double
                       dd_oo[ll1][ll2],int cc,int cctt1,double ppss)
    {  int f1,c1,d1,d3=-1;
       int cont11=0,cont15=0;
       double acum1=0,acum2=0,acum3=0,acum4=0,
            acum5=0,acum6=0,acum7=0,min_valor=maxint,max_valor=0.0;


         d1=-1;
        for (c1=0;c1<cctt1;c1++)
         { if (dd_oo[c1][0]==1) cont11++;

           if (dd_oo[c1][1]<min_valor){ min_valor=dd_oo[c1][1];
```

```
                    d1=c1;
                  };
  if (dd_oo[c1][1]>max_valor) max_valor=dd_oo[c1][1];
 }

f1=0;

if (d1!=-1) acum7=dd_oo[d1][1];


d1=0;
 do
 {
  cont15=0;
  acum5=0;
  acum6=0;
  acum4=0;
  acum2=0;


  {for (c1=0;c1<cctt1;c1++)
  { if  (dd_oo[c1][0]==cc)
    {
     acum2=(acum7- dd_oo[c1][1])/hh;
     acum4=c_kernel_Epa(acum2);
     acum5+=acum4;
     cont15++;

    }
   }

  }
   if ((cont15!=0))
 {
    acum1= double(cctt1-cont11);
    if (cc==0){ROC1[d1][0]=1-(acum5/acum1);

            }
    if (cc==1)ROC1[d1][1]=1-(acum5/double(cont11));


    ROC1[d1][2]=ceil(acum7);
    d1++;

  }
 acum7+=ppss;

}
```

```
        while ((acum7)< max_valor) ;

        }

double CV_kl(double dd_oo[ll1][ll2],double cctt,int cctt1)

      { int f1,c1,cont11=0,cont12=0;
        double acum1=0,acum2=0,acum3=0,acum4=0,acum5,valor_h=0,
            valor_act=0,v_fun=-maxint;
        time_t ini1,fin1;


        for (f1=0;f1<cctt1;f1++)
        {if (cctt==dd_oo[f1][0]) cont11++; }
        acum1=double(cont11-1);

        acum5= fabs((dd_oo[0][1]-dd_oo[cctt1-1][1])/cont11);


        system("cls");
        gotoxy(6,6);
        printf("Este algoritmo de calculo del Bandwidth  optimo esta basado en
              el \n");
        gotoxy(6,8);
        printf("metodo de la Maxima Verosimilitud, (validacion cruzada).\n");
        gotoxy(6,10);
        printf("Su computacion resulta costosa, pues la complejidad del
              algoritmo, \n");
        gotoxy(6,12);
        printf("es de orden  cinco veces n al cubo. \n");
        gotoxy(6,14);
        printf("Tardara unos minutos, no se
              impaciente.....COMPUTANDO.\n");
        ini1=time(NULL);
        do
        {
        valor_act+=acum5;

        acum3=0;
        for (f1=0;f1<(cctt1-1);f1++)
         { for (c1=0;c1<(cctt1-1);c1++)
           {
           if ((dd_oo[f1][1]!=dd_oo[c1][1])&& (dd_oo[f1][1]!=maxint)
             && (dd_oo[c1][1]!=maxint))

            {acum2=dd_oo[f1][1]-dd_oo[c1][1];

             acum2=acum2/valor_act;
```

```
      if (fabs(acum2)<=1)
      {
        {
        cont12++;

       acum2=0.75*(1-(pow(acum2,2)));

       acum3+=acum2;
        }
     if (fabs(acum2)>1) acum2=0;

      }

    if (acum3!=0) acum4=log(acum3);
    else acum4=-maxint;
  }
  }
  if (acum4!=-maxint) acum4=(acum4/(double(cont11)))-
                                    log(acum1*valor_act);

  if  (acum4 > v_fun)
    { valor_h=valor_act;
     v_fun=acum4;

    };
    }
  }
 while  (valor_act<dd_oo[cctt1-1][1]+ acum5);
  fin1=time(NULL);

  printf("                                          \n");
  printf("      El calculo se ha ejecutado en segundos %.1f \n",
              difftime(fin1,ini1));
  printf("                                          \n");
  if (cctt==0)  printf("      Valor del Bandwidth  Optimo para los Sanos:
                          %f\n", valor_h);
  if (cctt==1)  printf("      Valor del Bandwidth  Optimo para los
                          Enfermos:  %f\n", valor_h);
  printf("                                          \n");
  system("Pause");
  if ((cont11!=0)&& (v_fun>(-maxint))) return valor_h;
   else return 0;
}

double Media (double dd_oo[ll1][ll2],int cctt)
```

```
{ int f1,c1,cont11=0;
  double med=0;

   for (f1=0; f1<ll1; f1++)
    if (dd_oo[f1][0]==cctt)
        { cont11++;
          med+=dd_oo[f1][1];
        }
    med=med/cont11;

   return med;
}

double SD (double dd_oo[ll1][ll2],int cctt,double med1)
  { int f1,c1,cont11=0;
  double sd1=0;

   for (f1=0; f1<ll1; f1++)
    if (dd_oo[f1][0]==cctt)
        { cont11++;
          sd1+=pow(dd_oo[f1][1],2);
        }

    sd1=sd1/cont11;
    sd1=(sqrt(sd1 - pow(med1,2)));

    return sd1;

   }

double Pulgar (double dd_oo[ll1][ll2],int cctt,double zz)

 { double acum1,acum2,acum3,acum4;
  int f1,cont11=0;

    acum1=SD (dd_oo,cctt,Media(dd_oo,cctt));
    for (f1=0; f1<ll1; f1++)
     if (dd_oo[f1][0]==cctt)
         { cont11++;
         }
    acum2=floor(0.25*double(cont11));
    acum3=floor(0.75*double(cont11));
    acum2=fabs(acum3-acum2);
    acum4=acum2/acum1;

    acum3=acum2/zz;

    if (acum1<acum2)acum4=acum1;
```

```
    else acum4=acum2;

   return ((0.75*acum4)*(exp((-0.2*log(double(cont11))))));
 }

double Sondear_rel (double dd_oo[ll1][ll2],int cctt)

{ double acum1,acum2,acum3,acum4;
 int f1,cont11=0;

   acum1=SD (dd_oo,cctt,Media(dd_oo,cctt));
   for (f1=0; f1<ll1; f1++)
    if (dd_oo[f1][0]==cctt)
        { cont11++;
        }
   acum2=floor(0.25*double(cont11));
   acum3=floor(0.75*double(cont11));
   acum2=fabs(acum3-acum2);
   acum4=acum2/acum1;
   return acum4;
 }

void Estima_prob (double dd_oo[ll1][ll2],double ROC1[ll3][3],
            int ss,int dd,int lugar,int cctt1)

{int f1,c1,c3,c2,f2;
 double fp1,sensi2;

 f1=0;
 f2=0;
 c1=1;
 fp1=0;
 sensi2=0;

 do
 {
 if (ROC1[f2][2]!=0) f2++;
 else c1=0;
 }
 while ( c1==1);

 c1=0;c3=0;
 for (c2=0;c2<(cctt1-1);c2++) if (dd_oo[c2][0]==1) c1++;
 c3=cctt1-c1;


 switch (dd)
 {
```

```
case 0: // mujeres

   if  (ss==0) //Deplección.
   {
   for (f1=lugar;f1>-1;f1--)

    {
      if ( (dd_oo[f1][2]==0))
          { if (dd_oo[f1][0]==0) fp1++;  //FP
            else sensi2++;   //sensibilidad

          }
     }

     }
    else if ((ss==1)&& (dd_oo[f1][2]==0 ))//Elevación.
    { c1=ll1-lugar;

     for (f1=lugar;f1<ll1;f1++)
    {
      if ( (dd_oo[f1][2]==0))
          { if (dd_oo[f1][0]==0) fp1++;
            else sensi2++;
          }

     }

   fp1=fp1/double(c1);
   sensi2=sensi2/double(c1);
   ROC1[f2][0]=fp1;
   ROC1[f2][1]=sensi2;
   ROC1[f2][2]=dd_oo[lugar][1];
   break;

case 1:  //hombres

  if  (ss==0) //Deplección.
    {  c1=lugar+1;
    for (f1=lugar;f1>-1;f1--)
      {
      if ( (dd_oo[f1][2]==1))
          { if (dd_oo[f1][0]==0) fp1++;  //FP
            else sensi2++;   //sensibilidad
           }
      }

     }
    else if (ss==1) //Elevación.
```

```
{ c1=ll1-lugar;
 for (f1=lugar;f1<ll1;f1++)

{
 if ( (dd_oo[f1][2]==1))
      { if (dd_oo[f1][0]==0) fp1++;
        else sensi2++;
       }
 }

fp1=fp1/double(c1);
sensi2=sensi2/double(c1);
 ROC1[f2][0]=fp1;
 ROC1[f2][1]=sensi2;
 ROC1[f2][2]=dd_oo[lugar][1];
break;

case 2:  //Todos

 if  (ss==0) // Bajada.
{
for (f1=0;f1<cctt1;f1)

 {
  if ((dd_oo[f1][0]==0)&& (dd_oo[f1][2]<=dd_oo[lugar][1])) fp1++ ;
//FP
  if ((dd_oo[f1][0]==1)&& (dd_oo[f1][2]<=dd_oo[lugar][1])) sensi2++ ;
//sensibilidad

   }

  }
 else if (ss==1) //Elevación.
 {  for (f1=0;f1<cctt1;f1++)

 {  if ((dd_oo[f1][0]==0)&& (dd_oo[f1][1]>=dd_oo[lugar][1])) fp1++;
//FP
   if ((dd_oo[f1][0]==1)&& (dd_oo[f1][1]>=dd_oo[lugar][1])) sensi2++;

   }

fp1=fp1/double(c3);
sensi2=sensi2/double(c1);
if ((fp1!=0)&&(sensi2!=0)&&(dd_oo[lugar][1]!=0))
{ROC1[f2][0]=fp1;
 ROC1[f2][1]=sensi2;
 ROC1[f2][2]=dd_oo[lugar][1];
 }
```

```
        break;


      }
     default:
      printf("   El valor %d no es valido\n",dd);

      system ("Pause");
       break;
      }
     }
   }
    }


int Busca_cut_off (double ROC1[ll3][3])
 {int  f1,f2;
  double acum1;

  f1=0;f2=0;
  acum1=0;
   while (ROC1[f1][2]!=0)
    {

    if (((1-ROC1[f1][0])+ROC1[f1][1])>acum1)
       {
        f2=f1;
        acum1=(1-ROC1[f1][0])+ROC1[f1][1];
       }
    f1++;
    }

   return f2;

 }

int Busca_cut_off1 (double ROC1[ll3][3])
 {int  f1,f2;
  double acum1;

  f1=0;f2=0;
  acum1=0;
   while (ROC1[f1][2]!=0)
    {

    if ((((1-ROC1[f1][0])+ROC1[f1][1])>acum1)&& (ROC1[f1][1]>=0.9))
       {
```

```
            f2=f1;
            acum1=(1-ROC1[f1][0])+ROC1[f1][1];
           }
        f1++;
       }

      return f2;
      }

  void Grabar (double ROC[ll3][3], FILE *ffch1)
  { int f1,j1;

     f1=0;
     j1=0;
      do
      { if ((j1==1)&& (ROC[f1][2]!=ROC[f1-1][2]))
       {fprintf(ffch1,"%f %f %f \n",ROC[f1][0],ROC[f1][1],ROC[f1][2]);}
       if (j1==0) fprintf(ffch1,"%f %f %f
\n",ROC[f1][0],ROC[f1][1],ROC[f1][2]);
        j1=1;
        f1++;
       }
       while (ROC[f1][2]!=0);
   }

   void Alisa_curva (double ROC1[ll3][3],int ccont)
    {int ctr1=0,f3=0,f1=0,f2=0;


      for (f1=0;f1<ccont;f1++)
       { if ((ROC1[f1][2]==0) && (ROC1[f1][1]==0) && (ROC1[f1][0]==0))
          {
           for (f2=f1; f2<ccont ;f2++)
           { ROC1[f2][0]=ROC1[f2+1][0];
            ROC1[f2][1]=ROC1[f2+1][1];
            ROC1[f2][2]=ROC1[f2+1][2];
           }
          }

       }
      while ((ROC1[f1][2]!=0)&&(ROC1[f1+1][2]!=0))
      { ctr1=0;
      if ((fabs(ROC1[f1][0]- ROC1[f1+1][0])>0.2000)||
          (fabs(ROC1[f1][1]- ROC1[f1+1][1])>0.20000));
          {for (f2=f1; ROC1[f2][2]!=0;f2++)
            ROC1[f2][0]=ROC1[f2+1][0];
            ROC1[f2][1]=ROC1[f2+1][1];
            ROC1[f2][2]=ROC1[f2+1][2];
```

```
          ctr1=1;
        }

       f1++;
       }

    }

double Area_bajo_curva (double ROC1[ll3][3])
  {int f1;
   double acum1=0;
   f1=0;

   while ((ROC1[f1+1][2]!=0))
    { acum1+= ((ROC1[f1+1][1]+ROC1[f1][1])/2)*fabs(ROC1[f1+1][0]-
ROC1[f1][0]);
      f1++ ;
     }

   return acum1;
  }


void  Valor_pred_pos ( double ROC1[ll3][3],float prev1)
  { int f11=0,f21=0, para11=0;
    float auxv3,auxv13,auxv23,auxv33;
    double acum41,acum51;


     auxv33=prev1;
     auxv23=1-auxv33;
     acum51=0;
     while ((ROC1[f11][2]!=0) && (para11==0))
     {

     acum41=ROC1[f11][1];
     auxv3=float(acum41)*auxv33;
     acum41=ROC1[f11][0];


     auxv13=auxv23*float(acum41);
     auxv13+=auxv3;
     if (auxv13!=0) acum41=(double(auxv3))/(double(auxv13));


     if (acum41 >= 0.95) para11=1;
     if (acum41 > acum51)
       {acum51=acum41;
```

```
      f21=f11;
     }
    f11++;

   }
 if (para11==1)
 {ROC1[ll3][2]=ROC1[f11-1][2];
  if (acum41>1) acum41=1.0;
  ROC1[ll3][1]=acum41;
 }
 else
 { ROC1[ll3][2]=ROC1[f21][2];
   if (acum51>1) acum51=1.0;
   ROC1[ll3][1]=acum51;

 };

 }

void  Valor_pred_neg ( double ROC1[ll3][3],float prev1)
 { int f1=0,f2=0, para1=0;
   float auxv,auxv1,auxv2,auxv3;
   double acum4,acum5;


    auxv3=prev1;
    auxv2=1-auxv3;

    acum5=0;
    while ((ROC1[f1][2]!=0) && (para1==0))
    {

    acum4=ROC1[f1][0];

    acum4=1-acum4;

    auxv=float(acum4)*auxv2;

    acum4=ROC1[f1][1];
    acum4=1-acum4;
    auxv1=auxv3*float(acum4);
    auxv1+=auxv;
    if (auxv1!=0) auxv=float(auxv)/float(auxv1);

    acum4=auxv;

   if (acum4 >= 0.80) para1=1;
```

```
      f1++ ;
     }

  if (para1==1)
  {ROC1[ll3-1][2]=ROC1[f1-1][2];
   if (acum4>1) acum4=1.0;
   ROC1[ll3-1][1]=acum4;
  }
  else
  { ROC1[ll3-1][2]=ROC1[f2][2];
    if (acum5>1) acum5=1.0;
    ROC1[ll3-1][1]=acum5;
   }
 }

void Limpia (double ROC1[ll3][3])
   {int f1,c1;
    for (f1=0;f1<ll3;f1++)
     { for (c1=0;c1<3;c1++) ROC1[f1][c1]=0;}

     }

void Busca_en_tabla (int coor_nor1[2], double ddtt1, double mm1,double vv1)
    { double dif,difl,aux12;
     int f2;
     { for (f2=0;f2<3;f2++) coor_nor1[f2]=0;

       aux12=((ddtt1-mm1)/vv1);
       dif=floor(fabs(aux12));
       difl=floor((fabs(aux12)-dif)*10);
       aux12= (fabs(aux12)-dif)-difl*0.1;
       aux12=floor(aux12*100);
       coor_nor1[0]=int(10*dif)+difl;
       coor_nor1[1]=int(aux12);
     }
    }
void Calcula_prob_nor (int coor_nor1[2], double normal1[35][10],double ddtt1,
                       char kkaa, double ROC1[ll3][3],double med_var1[2][2])
    {double  dif,difl,aux11;
     int f1,f2,c1,c2,cont11;

      f1=0;
      f2=0;

      c2=0;

      cont11=0;
```

```
do
{
if (ROC1[f1][0]==0) cont11=1;
f1++;
}

while (cont11!=1) ;
cont11=f1-1;
ROC1[cont11][2]=ddtt1;


if (toupper(kkaa)=='S') // Aumento.
  { Busca_en_tabla (coor_nor1, ddtt1, med_var1[0][0],med_var1[1][0]);

  aux11=((ddtt1-med_var1[0][0])/med_var1[1][0]); //FP
  //printf(" valor de la puntuación %f \n",aux11);
   f2=coor_nor1[0];
   c2=coor_nor1[1];

   if (f2>34){if (aux11 < 0) ROC1[cont11][0]=0.9999;
        else ROC1[cont11][0]=0.0001; }

    else {if (aux11 < 0) ROC1[cont11][0]=normal1[f2][c2];
        else ROC1[cont11][0]=1-normal1[f2][c2];};

   Busca_en_tabla ( coor_nor1,ddtt1, med_var1[0][1],med_var1[1][1]);
   aux11=(ddtt1-med_var1[0][1])/med_var1[1][1]; //Sensibilidad
   //printf(" valor de la puntuación %f \n",aux11);

    f2=coor_nor1[0];
    c2=coor_nor1[1];


   if (f2>34){if (aux11 < 0) ROC1[cont11][1]=0.9999;
        else ROC1[cont11][1]=0.0001; }

    else {if (aux11 < 0) ROC1[cont11][1]=normal1[f2][c2];
        else ROC1[cont11][1]=1-normal1[f2][c2];};
   }

else { // Bajada

   aux11=((ddtt1-med_var1[0][0])/med_var1[1][0]); //FP
   Busca_en_tabla ( coor_nor1,ddtt1, med_var1[0][0],med_var1[1][0]);

    f2=coor_nor1[0];
    c2=coor_nor1[1];
```

```
            if (f2>34){if (aux11 < 0) ROC1[cont11][0]=0.9999;
                else ROC1[cont11][0]=0.0001; }

            else {if (aux11 < 0) ROC1[cont11][0]=1-normal1[f2][c2];
                else ROC1[cont11][0]=normal1[f2][c2];};



          aux11=(ddtt1-med_var1[0][1])/med_var1[1][1]; //Sensibilidad
          Busca_en_tabla (coor_nor1,ddtt1, med_var1[0][1],med_var1[1][1]);
          f2=coor_nor1[0];
          c2=coor_nor1[1];

           if (f2>34){if (aux11 < 0) ROC1[cont11][1]=0.0001;
                else ROC1[cont11][1]=0.9999; }

            else {if (aux11 < 0) ROC1[cont11][1]=1-normal1[f2][c2];
                else ROC1[cont11][1]=normal1[f2][c2];};

                 };
  }

 void copiar ( double dd_oo[ll1][ll2],double dd_oo_c[ll1][ll2])
  {int ff1,cc1;
    for (ff1=0;ff1<ll1;ff1++)
     {for (cc1=0;cc1<ll2;cc1++)
       dd_oo_c[ll1][ll2]=dd_oo[ll1][ll2];}

  }


main ()
{
FILE *fch1,*fch2,*f_normal,*fch3,*fch4,*pfs=NULL;
 double normal[35][10], datos_orig[ll1][ll2],datos_orig_c[ll1][ll2],
      ab_r,ab_n=-1,ab_np=-1,auxd,x,y,z,ROC[ll3][3],G1,G2,G3,
      med_var[2][2],G4,G5,G6,G7,G8,G9,G10,G11,G21,G31,G41,G51,G61,
      G71,G81,G91,G101;

 float vale0,vale1,vale2,vale3,vale4,vale5,vale6,vale7,vale8,vale9,prev,real;
 int ff,cc,aux0,aux1,aux2,aaux1,aaux2,auxm,cont,cont_v,cont1,
    coor_nor[2],cont_c,mar;
 char fi[40],ka,kb;
 time_t ini,fin;

 printf("  Este programa implementa el analisis de un protocolo o ");
 gotoxy(1,3);
 printf("  proceso diagnóstico, basandose en estimacion de curvas ROC.");
```

```
gotoxy(1,5);
printf("  Implementa tres variantes metodologicas: la estimacion cruda, ");
gotoxy(1,7);
printf("  otra basada en la distibucion normal y por ultimo la que utiliza ");
gotoxy(1,9);
printf("  metodos de estimacion no parametrica. En cada caso se calcula
        umbral,\n");
gotoxy(1,11);
printf("  intervalo y area bajo las curvas.
        \n");
gotoxy(1,13);
printf("  Es necesario un  fichero de datos de entrada, con resultados
        dignosticos.        \n");
gotoxy(1,15);
printf("                                                    \n");
gotoxy(1,20);
printf("    Para continuar pulse la tecla ENTER..... ");
getchar();


system("cls");
gotoxy(5,3);
printf("      ESTIMACION DE CURVAS ROC POR DIFERENTES
           METODOLOGIAS.  \n");
gotoxy(10,6);
printf(" Escriba el nombre del fichero de datos = ");
 //h1=fopen("simula500_roc.txt","r");
 //fch1=fopen("datos443_gluce.txt","r");
 //fch1=fopen("datos443_roc.txt","r");
gets(fi);
fflush(stdin);
gotoxy(10,9);
ka='S';
printf(" Debe indicar si la patologia es de Aumento. Responder (S/N)=  ");
scanf("%c",&ka);
fflush(stdin);

gotoxy(10,11);
printf(" Si quiere que la prevalencia de la enfermedad  la calcule el
        programa,\n");
gotoxy(10,12);
printf(" contestar S. Si la introduce manualmente (casos-controles),debe\n ");
gotoxy(10,13);
printf(" contestar N. (S/N)= " );
scanf("%c",&kb);
//printf(" valor de la pat %c", kb);
fflush(stdin);
prev=0;
```

```
if (toupper(kb)=='N')
  {gotoxy(10,15);
   printf(" Introduzca el valor de la prevalencia de la enfermedad en % =  " );
   scanf("%f",&prev);
   fflush(stdin);
   prev=0.01*prev;
   };

printf("                                    \n");
system("Pause");
fch1=fopen(fi,"r"); // "datos443_gluce.txt" y para edad "datos443_roc.txt"

if (fch1==NULL)
  {
 printf("  El fichero de datos no se puede abrir      \n");
 system("Pause");
 return -1;

  }
rewind(fch1);
fch2=fopen("Datos_ROC_real.TXT","w+");
   if (fch2==NULL)
    {
    printf("Error el fichero de grabacion, de datos reales, no se puede
            abrir.\n");
    system("Pause");
    return -1;
    }
 rewind(fch2);

   pfs=fopen("lpt1","w");
   if (pfs==NULL)
    {
    printf("Error el fichero de grabacion, de datos crudos, no se puede
            abrir.\n");
    system("Pause");
    return -1;
    }
//system("Pause");
 for (ff=0;ff<ll1;ff++)
   {for (cc=0;cc<4;cc++) datos_orig[ff][cc]=maxint;}

 ff=0;
 cc=0;
 cont=0;
 cont_v=0;
 cont_c=0;
 auxd=0;
```

```
while (!feof(fch1))
{
 fscanf(fch1,"%d %d %d ",&aux0,&aux1,&aux2);

 auxd=double(aux0);
 cc=0;
 datos_orig[ff][cc]=auxd;
 if (datos_orig[ff][0]==1) cont_c++ ;
 auxd=double(aux1);

 cc++;
 datos_orig[ff][cc]=auxd;

 auxd=double(aux2);

 cc++;
 datos_orig[ff][cc]=auxd;
  datos_orig[ff][3]=0;
  ff++;
  cont_v++;

 }
system("cls");
gotoxy(10,12);
printf(" El numero de individuos  es=   %d: Sanos %d y Enfermos
        %d\n",cont_v,cont_v-cont_c, cont_c);
 printf("                                    \n");

  do
  {cont=0;
  for (ff=0; ff<ll1-1; ff++)
   {

    if (datos_orig[ff][1]>datos_orig[ff+1][1])
       {aux0=datos_orig[ff+1][0];
        aux1=datos_orig[ff+1][1];
        aux2=datos_orig[ff+1][2];
        datos_orig[ff+1][0]=datos_orig[ff][0];
        datos_orig[ff+1][1]=datos_orig[ff][1];
        datos_orig[ff+1][2]=datos_orig[ff][2];
        datos_orig[ff][0]= aux0;
        datos_orig[ff][1]= aux1;
        datos_orig[ff][2]= aux2;
        cont=1;
        }
    };
  }
```

```
while(cont==1);
copiar ( datos_orig,datos_orig_c);


cont=1;
datos_orig[0][3]=1;
for (ff=1; ff<ll1-1; ff++)
{if ((datos_orig[ff][1]!=datos_orig[ff-1][1])&&
   (datos_orig[ff][1]!=maxint))
  { datos_orig[ff][3]=1;
    cont++ ;
   }
 }

printf("        Hay  %d  valores diferentes\n",cont);
printf("                                    \n");
system("Pause");

 if (prev==0)  prev=double(cont_c)/double(cont_v);

 for (ff=0;ff<ll1;ff++)
   {
  cont1=0;
  aux1=1;
  aux2=2;
  cc=0;

  if ((datos_orig[ff][3]==1)&& (datos_orig[ff][1]!=0))
  {
  Estima_prob (datos_orig,ROC,aux1,aux2,ff,cont_v);


  }
  aux1=1;

  cont1++;
   }

 Alisa_curva (ROC,cont);

ab_r=Area_bajo_curva (ROC);

 ff=0;

  Valor_pred_pos (ROC, prev);
  copiar ( datos_orig_c,datos_orig);

  Valor_pred_neg (ROC, prev);
```

```
    aux1=Busca_cut_off (ROC);

    aux2=Busca_cut_off1 (ROC);

  Grabar(ROC, fch2);
  G1=ROC[aux1][2];G2=ROC[aux1][1];G3=1-ROC[aux1][0];
  G4=ROC[aux2][2];G5=ROC[aux2][1];G6=1-ROC[aux2][0];
  G7= ROC[ll3][2]; G8=ROC[ll3][1] ;
  G9=ROC[ll3-1][2];G10=ROC[ll3-1][1];

  system("cls");
  gotoxy(5,3);
   printf("        ESTIMACION DE CURVAS ROC. R E S U L T A D O S.
\n");

printf("_____________________________________________________________
_____________________\n");

   printf("                                    \n");
   printf(" Numero de individuos procesados:  %4.0d\n",cont_v );
   printf("                                    \n");
   printf(" Numero de individuos enfermos     %4.0d\n",cont_c);
   printf("                                    \n");
   printf(" Numero de individuos sanos:       %4.0d\n",cont_v-cont_c );
   printf("                                    \n");
   printf(" La prevalencia de la enfermedad, usada en los calculos, es : %2.2f
         %\n",100*prev);
   printf("                                    \n");

printf("_____________________________________________________________
_____________________\n");

   printf("                                    \n");
   printf("                 METODO : ESTIMACION CRUDA.\n");
   printf("-------------------------------------------\n");
   printf(" CRITERIO MAX.(ESP.+SENS): Cut-off %2.2f  Sensi. %.4f     Esp.
         %.4f\n",ROC[aux1][2],ROC[aux1][1],1-ROC[aux1][0]);
   printf(" CRITERIO MAX. SENSI.   : Cut-off %2.2f  Sensi. %.4f     Esp.
         %.4f\n",ROC[aux2][2],ROC[aux2][1],1-ROC[aux2][0]);
   printf(" INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN %2.2f
         Prob. %.4f\n",ROC[ll3-1][2],ROC[ll3-1][1]);
   printf(" INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP %2.2f
         Prob. %.4f\n",ROC[ll3][2],ROC[ll3][1]);
   printf(" AREA BAJO LA CURVA ROC :       %.5f\n",    ab_r);
   printf("-------------------------------------------\n");

   printf(" Quiere realizar otra estimacion basada en el supuesto de normalidad
         ? (S/N)= ");
```

```
scanf("%c",&ka);

fflush(stdin);
 if (toupper(ka)=='S')
  {f_normal=fopen("tabla_normal.txt","r");

  if (f_normal==NULL)
   {
 printf("  El fichero de la tabla de la distribución normal no se puede abrir.
          \n");
  system("Pause");
  return -1;

  }
   rewind(f_normal);
   tabla_normal(normal,f_normal);
   fch3=fopen("Datos_ROC_nor.TXT","w+");
   if (fch3==NULL)
    {
    printf("Error el fichero de grabacion,de datos de la normal,no se puede
          abrir.\n");
    system("Pause");
    return -1;
    }
    rewind(fch3);

    Limpia (ROC);

    med_var[0][0]= Media (datos_orig,0);
    med_var[1][0]= SD (datos_orig,0,med_var[0][0]);
    med_var[0][1]= Media (datos_orig,1);
    med_var[1][1]= SD (datos_orig,1,med_var[0][1]);

     for (ff=1;ff<ll1-1;ff++)
   {
    if (datos_orig[ff][3]==1)
    { auxd=datos_orig[ff][1];
     coor_nor[0]=0;
     coor_nor[1]=0;

     Calcula_prob_nor(coor_nor,normal,auxd, ka, ROC,med_var);

    }

}

Alisa_curva (ROC,cont);
```

```
     ff=0;

     Valor_pred_pos (ROC, prev);
     Valor_pred_neg (ROC, prev);

     aux1=Busca_cut_off (ROC);

     aux2=Busca_cut_off1 (ROC);

     for (ff=0; ff<ll3; ff++)
      {
     if (ROC[ff][0]!=0)
     {//printf(" valores de Fp %f y Sensi. %f para la edad
%f\n",ROC[ff][0],ROC[ff][1],ROC[ff][2]);
    // printf("                                    \n");
     //getchar();
     }
     }

  ab_n= Area_bajo_curva (ROC);

  Grabar(ROC, fch3);
  G11=ROC[aux1][2];G21=ROC[aux1][1];G31=1-ROC[aux1][0];
  G41=ROC[aux2][2];G51=ROC[aux2][1];G61=1-ROC[aux2][0];
  G71= ROC[ll3][2];G81=ROC[ll3][1] ;
  G91=ROC[ll3-1][2];G101=ROC[ll3-1][1];


   system("cls");
   gotoxy(5,3);
   printf("        ESTIMACION DE CURVAS ROC. R E S U L T A D O S.
\n");

printf("_____________________________________________________________
_____________________\n");
    printf("                                    \n");
   printf(" Numero de individuos procesados:  %4.0d\n",cont_v );
   printf("                                   \n");
   printf(" Numero de individuos enfermos     %4.0d\n",cont_c);
   printf("                                   \n");
   printf(" Numero de individuos sanos:       %4.0d\n",cont_v-cont_c );
   printf("                                   \n");
   printf(" La prevalencia de la enfermedad, usada en los calculos, es : %2.2f
          %\n",100*prev);
   printf("                                   \n");

printf("_____________________________________________________________
_____________________\n");
```

```
printf("              METODO : ESTIMACION CRUDA.\n");
printf("-------------------------------------------\n");
printf(" CRITERIO MAX.(ESP.+SENS): Cut-off %2.2f  Sensi. %.4f    Esp.
        %.4f\n",G1,G2,G3);
printf(" CRITERIO MAX. SENSI.   : Cut-off %2.2f  Sensi. %.4f    Esp.
        %.4f\n",G4,G5,G6);
printf(" INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN %2.2f
        Prob. %.4f\n",G9,G10);
printf(" INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP %2.2f
        Prob. %.4f\n",G7,G8);
printf(" AREA BAJO LA CURVA ROC :       %.5f\n",   ab_r);
printf("-------------------------------------------\n");


printf("_____________________________________________________________
_____\n");

printf("              METODO : ESTIMACION NORMAL.\n");
printf("-------------------------------------------\n");
printf(" CRITERIO MAX.(ESP.+SENS): Cut-off %2.2f  Sensi. %.4f    Esp.
        %.4f\n",G11,G21,G31);
printf(" CRITERIO MAX. SENSI.   : Cut-off %2.2f  Sensi. %.4f    Esp.
        %.4f\n",G41,G51,G61);
printf(" INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN %2.2f
        Prob. %.4f\n",G91,G101);
printf(" INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP %2.2f
        Prob. %.4f\n",G71,G81);
printf(" AREA BAJO LA CURVA ROC :       %.5f\n",   ab_n);
printf("-------------------------------------------\n");
//printf("                                          \n");
system("Pause");


fclose(f_normal);
fclose(fch3);

   };
printf(" Quiere realizar otra estimacion basada en metodos no parametricos
       ? (S/N)= ");
scanf("%c",&kb);
//printf(" valor de la pat %c", kb);
fflush(stdin);

if (toupper(kb)=='S')
   {
   fch4=fopen("Datos_ROC_nopar.TXT","w+");
   if (fch4==NULL)
```

```
    {
    printf("Error el fichero de grabacion de datos no parametricos no se
            puede abrir.\n");
    system("Pause");
    return -1;
    }
    rewind(fch4);
    Limpia (ROC);

    system("cls");
 printf("                                        \n");
 printf("                                        \n");
 printf("                                        \n");
 printf("                                        \n");
 printf(" Para calcular el bandwidh optimo, se han implementado dos
         metodos: \n ");
 printf("                                        \n");
 printf(" La Regla del Pulgar, rapido, pero poco afinado; y otro metodo de
         Validacion \n ");
 printf("                                        \n");
 printf(" Cruzada,mejor, pero mas lento. Para ejecutar el primero escriba S,
         para el \n");
 printf("                                        \n");
 printf(" segundo pulse otra tecla ================> ");
 scanf("%c",&kb);

if (toupper(kb)=='S')
   {
  if (Sondear_rel (datos_orig,1)< Sondear_rel (datos_orig,0)) z=Sondear_rel
                                                    datos_orig,0);
    else z=Sondear_rel (datos_orig,1);
   x=Pulgar (datos_orig,1,z);//enfermos
   y=Pulgar (datos_orig,0,z);//sanos

    }

if (toupper(kb)!='S')
   {
    x=CV_kl(datos_orig,1,cont_v);
  if (x==0){ system("cls");
           gotoxy(1,10);
          printf("No puede calcular el parametro de alisamiento para los
                 casos. Introduzcalo a mano = ");
           scanf("%f",&x);
           fflush(stdin);
           };

  y=CV_kl(datos_orig,0,cont_v);
```

```
      if (y==0){ system("cls");
              gotoxy(1,14);
              printf("No puede calcular el parmetro de alisamiento para los
                    controles. Introduzcalo a mano = ");
              scanf("%f",&y);
              fflush(stdin);
              };
       }

       fflush(stdin);
      if (x>y) z=y;
       else z=x;
      k_integrado_Epa (ROC,y,datos_orig,0,cont_v,z);

      k_integrado_Epa (ROC,x,datos_orig,1,cont_v,z);
       cont=0;

       cont1=0;

       Valor_pred_pos (ROC, prev);
       Valor_pred_neg (ROC, prev);
       aux1=Busca_cut_off (ROC);

       aux2=Busca_cut_off1 (ROC);
       ab_np= Area_bajo_curva (ROC);
       ff=0;

       Grabar(ROC,fch4);

       system("cls");
   gotoxy(5,3);
   printf("        ESTIMACION DE CURVAS ROC. R E S U L T A D O S.
\n");

printf("_____________________________________________________________
_____\n");
   printf("                                    \n");
   printf(" Numero de individuos procesados:  %4.0d\n",cont_v );
   printf("                                    \n");
   printf(" Numero de individuos enfermos     %4.0d\n",cont_c);
   printf("                                    \n");
   printf(" Numero de individuos sanos:       %4.0d\n",cont_v-cont_c );
   printf("                                    \n");
   printf(" La prevalencia de la enfermedad, usada en los calculos, es : %2.2f
          %\n",100*prev);
   printf("                                    \n");
```

```
printf("_____________________________________________________________
___________________\n");
  printf("              METODO : ESTIMACION CRUDA.\n");
  printf("- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");
  printf(" CRITERIO MAX.(ESP.+SENS):  Cut-off %2.2f  Sensi. %.4f    Esp.
         %.4f\n",G1,G2,G3);
  printf(" CRITERIO MAX. SENSI.    : Cut-off %2.2f  Sensi. %.4f    Esp.
         %.4f\n",G4,G5,G6);
  printf(" INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN %2.2f
         Prob. %.4f\n",G9,G10);
  printf(" INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP %2.2f
         Prob. %.4f\n",G7,G8);
  printf(" AREA BAJO LA CURVA ROC  :        %.5f\n",    ab_r);
  printf("- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");

printf("_____________________________________________________________
___________________\n");
  printf("              METODO : ESTIMACION NO PARAMETRICA.\n");
  printf("- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");
  printf(" CRITERIO MAX.(ESP.+SENS):  Cut-off %2.2f  Sensi. %.4f    Esp.
         %.4f\n",ROC[aux1][2],ROC[aux1][1],1-ROC[aux1][0]);
  printf(" CRITERIO MAX. SENSI.    : Cut-off %2.2f  Sensi. %.4f    Esp.
         %.4f\n",ROC[aux2][2],ROC[aux2][1],1-ROC[aux2][0]);
  printf(" INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN %2.2f
         Prob. %.4f\n",ROC[ll3-1][2],ROC[ll3-1][1]);
  printf(" INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP %2.2f
         Prob. %.4f\n",ROC[ll3][2],ROC[ll3][1]);
  printf(" AREA BAJO LA CURVA ROC  :        %.5f\n",    ab_np);
  printf("- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");
  //printf("                                        \n");
  printf(" Parametros de Alisamiento:  Enfermos= %3.3f  y  Sanos=
        %3.3f\n",x,y);
  //printf("                                        \n");
  printf("- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");

printf("_____________________________________________________________
___________________\n");

  fclose(fch4);
     };
   printf("                                        \n");

   printf("                                        \n");
   printf(" Quiere imprimer los resultados ? (S/N)= ");
   scanf("%c",&kb);
     fflush(stdin);
```

```
if (toupper(kb)=='S')
{
fprintf(pfs,"        ESTIMACION DE CURVAS ROC. R E S U L T A D O S.
\n");

fprintf(pfs,"____________________________________________________
____________________\n");
fprintf(pfs,"                              \n");
fprintf(pfs," Numero de individuos procesados:  %4.0d\n",cont_v );
fprintf(pfs,"                              \n");
fprintf(pfs," Numero de individuos enfermos     %4.0d\n",cont_c);
fprintf(pfs,"                              \n");
fprintf(pfs," Numero de individuos sanos:      %4.0d\n",cont_v-cont_c );
fprintf(pfs,"                              \n");
fprintf(pfs," La prevalencia de la enfermedad, usada en los calculos, es :
          %2.2f %\n",100*prev);
fprintf(pfs,"                              \n");

fprintf(pfs,"____________________________________________________
____________________\n");

fprintf(pfs,"             METODO : ESTIMACION CRUDA.\n");
fprintf(pfs,"- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");
fprintf(pfs," CRITERIO MAX.(ESP.+SENS):  Cut-off %2.2f  Sensi. %.4f
          Esp. %.4f\n",G1,G2,G3);
fprintf(pfs," CRITERIO MAX. SENSI.    : Cut-off %2.2f  Sensi. %.4f    Esp.
          %.4f\n",G4,G5,G6);
fprintf(pfs," INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN
          %2.2f Prob. %.4f\n",G9,G10);
fprintf(pfs," INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP
          %2.2f Prob. %.4f\n",G7,G8);
fprintf(pfs," AREA BAJO LA CURVA ROC  :      %.5f\n",   ab_r);
fprintf(pfs,"- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");
if (ab_n!=-1)
  {
fprintf(pfs,"____________________________________________________
____________________\n");

fprintf(pfs,"             METODO : ESTIMACION NORMAL.\n");
fprintf(pfs,"- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");
fprintf(pfs," CRITERIO MAX.(ESP.+SENS):  Cut-off %2.2f  Sensi. %.4f
          Esp. %.4f\n",G11,G21,G31);
fprintf(pfs," CRITERIO MAX. SENSI.    : Cut-off %2.2f  Sensi. %.4f    Esp.
          %.4f\n",G41,G51,G61);
fprintf(pfs," INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN
          %2.2f Prob. %.4f\n",G91,G101);
fprintf(pfs," INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP
          %2.2f Prob. %.4f\n",G71,G81);
```

```
    fprintf(pfs," AREA BAJO LA CURVA ROC  :        %.5f\n",     ab_n);
    fprintf(pfs,"- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");

     }
     if (ab_np!=-1)
     {
fprintf(pfs,"___________________________________________________________________________
______________________\n");

    fprintf(pfs,"                METODO : ESTIMACION NO
PARAMETRICA.\n");
    fprintf(pfs,"- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - \n");
    fprintf(pfs," CRITERIO MAX.(ESP.+SENS):  Cut-off %2.2f  Sensi. %.4f
                Esp. %.4f\n",ROC[aux1][2],ROC[aux1][1],1-ROC[aux1][0]);
    fprintf(pfs," CRITERIO MAX. SENSI.   : Cut-off %2.2f  Sensi. %.4f     Esp.
                %.4f\n",ROC[aux2][2],ROC[aux2][1],1-ROC[aux2][0]);
    fprintf(pfs," INTERVALO basado en Valor Pred.Pos.y Neg:  Cut-off VPN
                %2.2f Prob. %.4f\n",ROC[ll3-1][2],ROC[ll3-1][1]);
    fprintf(pfs," INTERVALO basado en Valor Pred.Pos y Neg:  Cut-off VPP
                 %2.2f Prob. %.4f\n",ROC[ll3][2],ROC[ll3][1]);
    fprintf(pfs," AREA BAJO LA CURVA ROC  :        %.5f\n",     ab_np);
    fprintf(pfs," - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -\n");
    fprintf(pfs,"                                        \n");
    fprintf(pfs," Parametros de Alisamiento:  Enfermos= %3.3f  y  Sanos=
                 %3.3f\n",x,y);
    fprintf(pfs,"                                        \n");

fprintf(pfs,"___________________________________________________________________________
______________________\n");

     }
    fprintf(pfs,"                                        \n");

    }
      // Final del programa.
   system("cls");
   gotoxy(20,20);
   printf ("Programa realizado por E. Nunnez (2003.)\n");
   _sleep(2);
   system("cls");
   fclose(fch1);
   fclose(fch2);
   fclose(pfs);
}


 void tabla_normal (double tt_nn[35][10],FILE *f_normal)
```

```
{  int ff1,c1,cont11;

   float aux0,aux1,aux2,aux3,aux4,aux5,aux6,aux7,aux8,aux9;

    ff1=0;
    cont11=0;
   while (!feof(f_normal))
  {


  fscanf(f_normal,"%f %f %f %f %f %f %f %f %f %f
",&aux0,&aux1,&aux2,&aux3,&aux4,&aux5,&aux6,&aux7,&aux8,&aux9);

    c1=0;
     tt_nn[ff1][c1]=aux0;

   c1++;
   tt_nn[ff1][c1]=aux1;


   c1++;
   tt_nn[ff1][c1]=aux2;

   c1++;
   tt_nn[ff1][c1]=aux3;

   c1++;
   tt_nn[ff1][c1]=aux4;
   c1++;
   tt_nn[ff1][c1]=aux5;
   c1++;
   tt_nn[ff1][c1]=aux6;
   c1++;
   tt_nn[ff1][c1]=aux7;
   c1++;
   tt_nn[ff1][c1]=aux8;
   c1++;
   tt_nn[ff1][c1]=aux9;

    ff1++;
    cont11++;
    }
   system("cls");

   }
//-------------------------------------------------------------------------//
```

I.4. Factor de corrección lineal.

Este programa implementa un algoritmo de corrección para el vector beta, que contiene los coeficientes de regresión logística. El algoritmo fue revisado por F. Harrell, y se publico en la revista *Statictics in Medicine*. Esta aplicación usa como apoyo la biblioteca de funciones NAG (*Numerical Algoritm Group*) Mark 6 . El programa necesita un fichero de datos de entrada, compuesto por los valores de casos y controles.

Se ha planteado, para el cálculo del Factor de Corrección final, una alternativa consistente en eliminar los valores mínimos. Figura como FC corregido.

El algoritmo consiste en generar, mediante metodología bootstrap y usando como base el método de montecarlo, datos para generación de nuevos datos, sobre las que calcularemos nuevos factores de corrección. El programa produce dos FC, el de Harrel puro, y la corrección del mismo ya comentada.

En pasos iniciales e intermedios figuran la estimación del modelo real, cálculos intermedios de factores de corrección, índices *IP* y *Deviance* para cada modelo.

COMPLEJIDAD: Si llamamos *n* a la longitud del fichero de entrada, y *k* número de generaciones bootstrap; la complejidad del algoritmo es aproximadamente del orden de *k*n*; esto es, proporcional al valor de *n*; lineal ====➔ O(*n*).

Código fuente:

```
/*
 * Programa que calcula un factor de corrección del vector beta.
 * para regresión logística bidimensional; basado en el algoritmo
 * de E.W. Steyerberg,F. Harrell y otros.
 * Usa librerias NAG Mark6
 *
 *
 * E. Nuñez (2002).
 */

#include <vcl.h>
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
```

```
#include <iostream.h>
#include <fstream.h> //flujo de archivos
#include <conio.h>  //coordenadas de pantalla
#include <memory.h> // matrices dinámicas.
#include <time.h>
#include <dos.h>
namespace NAGspace {
#include <nag.h>
#include <nagg02.h>
#include <nag_stdlib.h>
}

const long m1=9;
const long n1=48;
const long ip=3;
const long ip1=1;
const long m2=1;

 void  montecarlo (int cont11,double datos_orig1[n1][m1],double
  datos_nag1[n1][m1],double yy_bb[n1],double yy[n1])
  {
   int ccont,i,j;
    for (i=0; i<cont11; i++)
    {
     ccont =  rand() % cont11;
     yy_bb[i]=yy[ccont];

    for (j=0; j<9; j++) datos_nag1[i][j] = datos_orig1[ccont][j];
        }
     }

 void calcula_IP ( double tt_dd[m1+1],double dd_oo[n1][m1],double
                   ii_pp[n1])
  {int fil1,col1;

   for (fil1=0;fil1<n1;fil1++)ii_pp[fil1]=0;

   for (fil1=0;fil1<n1;fil1++)
       {
      for (col1=0;col1<m1;col1++)
       { if (ii_pp[fil1]==0) ii_pp[fil1]+= tt_dd[col1+1]*dd_oo[fil1][col1];

          }
       /* Parte opcional de visualización de datos.

       //if (((fil1==ceil(n1/2))&&
      ((cconnt==ceil((10*BB)/2))||(cconnt==0)||(cconnt==10*BB))))
       //{printf("                                    \n");
```

```
                //printf("Datos IP bootstrap en fila %d %f\n",fil1,ii_pp[fil1]);
                //system("Pause");
                }
                 }
               }

          //Parte  especifica de funciones NAG
          using  namespace NAGspace;
           static int set_enum(char linkc, Nag_Link *link, char meanc,
                      Nag_IncludeMean *mean) ;
         main ()

         {

           NAGspace::Integer rank, sx[m1]={1,0,0,0,0,0,0,2,0},sx1[1]={1};;
           NAGspace::Integer print_iter=1000;
           NAGspace::Integer m,n;
           int auxdi[10];
           int ff,cc,fil,col, B,cont1,cont=0,aux0,aux1,aux2,aux3,aux4,
             aux5,aux6,aux7,aux8,aux9,max_iter;
          Nag_Link  link;
          Nag_IncludeMean  mean;

           FILE *f1,*f2;
           double
datos_orig[n1][m1],datos_nag[n1][m1],tdatos_nag[m1+1],tdatos_ori[m1+1],
se_aux1[m1+1],se_err[m2],se_aux2[m1+1],i_p[n1],i_p_ori[n1],i_p_com[n1],
                vbeta_max[m1],*mat=NULL,error_1;

           float vale0,vale1,vale2,vale3,vale4,vale5,vale6,vale7,vale8,vale9;
           double
cov[ip*(ip+1)/2],cov_err[ip1*(ip1+1)/2],se[ip],vbeta_boots[ip],vbeta_err[ip1],de
v,dev_ori,dev_err,df,*offset=(double *)0,auxd,aux,tol,eps,
                *wt=NULL, y[n1], y_b[n1], binom_t[n1] ,v[n1][ip+6]
                 ,v_err[n1][ip1+6] ,vbeta_ori[ip],vbeta_aux[ip];

          static  NagError fail;
          char buf[100],llave,fi[40];
          time_t ini,fin;

          // Inicializa el generador de números aleatorios
          time_t t;
          srand((unsigned) time(&t));

          printf("  Este programa implementa un algoritmo de corrección para
                    el vector beta ");
          gotoxy(1,3);
          printf("  que contiene los coeficientes de  Regresión Logística  .");
```

```
gotoxy(1,5);
printf(" El algoritmo fue revisado por Steyerberg, y se publico en la
        revista ");
gotoxy(1,7);
printf(" Statictics in Medicine. Esta aplicación usa como apoyo la
       biblioteca ");
gotoxy(1,9);
printf(" de funciones NAG. El programa necesita un fichero de
       datos de entrada, \n");
gotoxy(1,11);
printf(" compuesto por casos y controles.
         \n");
gotoxy(1,13);
printf(" Se ha planteado, para el calculo del Factor de Corrección
        final, una alternativa,\n");
gotoxy(1,15);
printf(" consistente en eliminar los valores minimos. Figura como
       FC corregido. \n");
gotoxy(1,20);
printf("  Para continuar pulse la tecla ENTER..... ");
getchar();
gotoxy(5,23);
printf(" Escriba el nombre del fichero de datos y pulse ENTER = ");

gets(fi);
fflush(stdin);
f1=fopen(fi,"r");   //f1=fopen("datos48.txt","r");
system("cls");
if (f1==NULL)
  {
  wsprintf(buf, "Imposible abrir fichero de entrada de datos");
  MessageBox(0, buf, "", MB_OK);
  return -1;
  }
rewind(f1);

for(fil=0;fil<n1;fil++){ binom_t[fil]=1;
                //printf("valor binom %f ", binom_t[fil]);
                //system("Pause");
                }
 ff=0;
 cc=0;

// Bucle de lectura del fichero de entrada.

while (!feof(f1))
  {
  fscanf(f1,"%d %d %d %d %d %d %d %d %d
```

```
        %d",&aux0,&aux1,&aux2,&aux3,&aux4,
         &aux5,&aux6,&aux7,&aux8,&aux9);
printf("                                    \n");
printf("                                    \n");
printf("                                    \n");
auxd=aux1;
datos_orig[ff][cc]=auxd;
auxd=aux2;

cc++;
datos_orig[ff][cc]=auxd;
auxd=aux3;

cc++;
datos_orig[ff][cc]=auxd;
auxd=aux4;

cc++;
datos_orig[ff][cc]=auxd;
auxd=aux5;

cc++;
datos_orig[ff][cc]=auxd;
auxd=aux6;

cc++;
datos_orig[ff][cc]=auxd;
auxd=aux7;
cc++;
datos_orig[ff][cc]=auxd;
auxd=aux8;
cc++;
datos_orig[ff][cc]=auxd;
auxd=aux9;
cc++;
datos_orig[ff][cc]=auxd;

 cc=0;
 auxd=aux0;
 y[ff]=auxd;

 cont++;
    ff++;

 /* Bucle opcional de comprobación de carga de datos.

     for (ff=0;ff<496;ff++)
      {for (cc=0;cc<9;cc++)
```

```
      {printf( "valores %f %f  \n",y[ff],datos_orig[ff][cc]);

      printf("                                        \n");
      getchar();
            }
            }*/

}
gotoxy(4,20);
printf("Se han cargado los datos de %d individuos\n",cont);
printf("                                        \n");
system("Pause");
system("cls");
gotoxy(2,15);
printf("ESCRIBA EL NUMERO DE GENERACIONES
        BOOTSTRAP(x10) QUE VA A REALIZAR===>");
scanf("%d",&B);
 ini=time(NULL);
 // Se comprueba que variables de la base son continuas o
   discretas
 /* for (fil=0;fil<m1;fil++) if sx[fil]!=0
                          {
                          cont=0;
                          cont1=0;
                          llave=0;
                          do
                          {
               if datos_orig[1][fil]!=datos_orig[cont][fil] cont1++;
               if (cont1>4) {llave=1; sx[fil]=2;};
                          cont++ ;
                          }
                          while ((cont<n1) && (llave==0))

                          } */

 cont1=0;
 cont=0;

//A continuación se crea una matriz dinámica para los FC's.

 aux4=10*B;
 aux4=aux4*sizeof(double);
 if((mat=(double *)malloc(aux4))==NULL)
  {
   printf("Insuficiente espacio en memoria\n");
   return -1;
  }
```

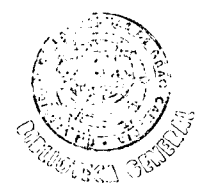
```
memset(mat,0,aux4);
//printf("La memoria dinamica se ha asignado \n");
f2=fopen("bootstrap.TXT","w+");
  if (f2==NULL)
   {
   printf("Error el fichero bootstrap no se puede abrir, pulse Enter
         para salir\n");
   exit(1);
   }

   /* Parámetros complementario para la función Nag g02gbc, que
      calcula el vector beta de RL.  */
   max_iter=100;
   tol=5e-5;
   eps=1e-6;
   aux4=ip+6;
   link=Nag_Logistic;
   mean=Nag_MeanInclude;
   aux7=n;
   aux8=m;
   aux9=ip;

   /* A continuacion viene la llamada a la función NAG g02gbc que
      realiza  el cálculo del vector beta  basado en los datos
      originales. Toma, como parámetro,  las variables incluidas en
      el modelo, pues no realiza la selección */

  g02gbc(link,mean,n1, (double *)datos_orig,m1,m1,sx,ip,y,binom_t,
  wt,offset,&dev_ori,&df,vbeta_ori,&rank,se,cov,(double *)v,

(NAGspace::Integer)(ip+6),tol,max_iter,print_iter,"",eps,&fail);//NA
GERR_DEFAULT

// Diversos mensajes, según código de fallo.

   if (fail.code == NE_NOERROR || fail.code ==
                  NE_SVD_NOT_CONV ||
    fail.code == NE_LSQ_ITER_NOT_CONV ||
    fail.code == NE_RANK_CHANGED || fail.code ==
                 NE_ZERO_DOF_ERROR)
   { system("cls");
    gotoxy (10,5);
    Vfprintf(stdout,"        ESTIMACION DEL MODELO
                          ORIGINAL\n\n");
    Vfprintf(stdout,"          Deviance = %12.4e\n", dev_ori);
    Vfprintf(stdout,"          Grados de libertad = %3.1f\n\n", df);
    Vfprintf(stdout,"        Estimacion del vector beta y su
                          Standard error\n\n");
```

```
      for (fil=0; fil<ip; fil++)
      Vfprintf(stdout,"          %14.4f%14.4f\n", vbeta_ori[fil], se[fil]);
      system("Pause");
      Vfprintf(stdout,"\n");
      /* Esta opción es de fallo y solo se activará cuando el
         programa aborte.
      Vfprintf(stdout,"   binom_t   y  fitted value  Residual
                          Leverage\n\n");
      for (fil = 0; fil < n1; ++fil)
       {
        Vfprintf(stdout,"%10.1f%7.1f%10.2f%12.4f%10.3f\n",
                 binom_t[fil], y[fil],
              v[fil][1], v[fil][4], v[fil][5]);
              //system("Pause");
       }
     }
    else
     {
      Vfprintf(stdout,"%s\n",fail.message);
       Vfprintf(stdout,"Uno o ambos  m1 y n1 podrían estar fuera de
                        rango:\
                  m1 = %-3ld mientras  n1 = %-3ld\n", m1, n1);
      fclose(f1);
      fclose(f2);

return(-1);

 //Fin de la opción de fallo */
    };


  if (mean==Nag_MeanInclude) {tdatos_ori[0]=vbeta_ori[0];
                      se_aux1[0]=se[0];}

   cont=1;
    for (fil=0;fil<m1;fil++)
          if (sx[fil]!=0)
          { tdatos_ori[fil+1]=vbeta_ori[cont];
           se_aux1[fil+1]=se[cont];
          cont++;
          };

   calcula_IP (tdatos_ori ,datos_orig,i_p_ori);

    for (fil=0;fil<m1;fil++) vbeta_max[fil]=-100;

             cont1=0;
    // Bucle principal del programa.
```

```
do
{

 llave=1;
 if (cont1>1) for (fil=0;fil<ip;fil++) vbeta_aux[fil]=vbeta_boots[fil];

  montecarlo (n1,datos_orig,datos_nag,y_b,y);


 /* A continuación viene la llamada a la función NAG g02gbc que
    realiza  el cálculo del vector beta, basado en generación
     bootstrap, realizadas por la función anterior "montecarlo". */

  mean=Nag_MeanInclude;
g02gbc(link,mean,n1,(double*)datos_nag,m1,m1,sx,ip,y_b,binom_t,
       wt,offset,&dev,&df,vbeta_boots,&rank,se,cov,(double*)v,
       (NAGspace::Integer)(ip+6),tol,max_iter,print_iter,"",eps,
       NAGERR_DEFAULT);

      cont=0;

        do
        {
          if (vbeta_aux[fil]!=vbeta_boots[fil]) llave=0;
          cont++;
         }
        while ((cont<ip)&& (llave==0)) ;

  if (llave==0)
  { cont=1;

  if (mean==Nag_MeanInclude) {tdatos_nag[0]=vbeta_boots[0];
                     se_aux2[0]=se[0];}
  for (fil=0;fil<m1;fil++)
       if (sx[fil]!=0)
       { tdatos_nag[fil+1]=vbeta_boots[cont];
         se_aux2[fil+1]=se[cont];

         cont++;
        };

   vale8=0;
   vale9=0;
   calcula_IP (tdatos_nag ,datos_orig,i_p);

    mean=Nag_MeanZero;
    g02gbc(link,mean,n1,
           (double *)i_p,m2,m2,sx1,ip1,y_b,binom_t,wt,offset,
```

```
                &dev_err,&df,vbeta_err,&rank,se_err,cov_err,
                (double *)v_err,(NAGspace::Integer)(ip1+6),tol,
                max_iter,print_iter,"",eps,NAGERR_DEFAULT);
    do
     {
     g02gbc(link,mean,n1,(double *)i_p,m2,m2,sx1,ip1,y_b,binom_t,
              wt,offset,&dev_err,&df,vbeta_err,&rank,se_err,cov_err,
             (double *)v_err,(NAGspace::Integer)(ip1+6),tol,max_iter,
             print_iter,"",eps,NAGERR_DEFAULT);

    printf(" Factor de correcion parcial %f =\n", vbeta_err[1]);
     system("Pause");

if (vbeta_err[1]<0.95) for (fil=0;fil<n1;fil++){ printf(" valor del  ip %f
                                                    \n",i_p[fil]);
 i_p[fil]=vbeta_err[1]*i_p[fil];
 printf(" valor del nuevo ip %f \n",i_p[fil]);
     }                                  }

 while  (vbeta_err[1] < 0.95); */
mat[cont1]=vbeta_err[1];
vale6=0;
vale7=0;

/* Parte de comprobación, es opcional.

for (fil=0;fil<m1;fil++)
          {
           if ((vbeta_max[fil]>tdatos_nag[fil+1]) &&
             (vbeta_max[fil]!=-100))
           { if ((sx[fil]==1) && ((exp(tdatos_nag[fil+1]))<1))
                {if (tdatos_nag[fil+1]>0)


                };
             if ((sx[fil]==2) && (exp(tdatos_nag[fil+1])<1.3))
                {

                };
           };

           if (vbeta_max[fil]<tdatos_nag[fil+1])
             vbeta_max[fil]=tdatos_nag[fil+1];

          }; */

         if ((cont1==ceil((10*B)/2))||(cont1==0)||(cont1==10*B))
        { aux1=ceil( n1/2);
```

```
    system("cls");
    printf("                                                    \n");
    if (cont1==0) printf("              GENERACION
                        BOOTSTRAP 1 RESULTADOS.  \n");
     else printf("              GENERACION BOOTSTRAP %d
                RESULTADOS.  \n",cont1);
    printf("                                          \n");
    printf("Componentes del vector Beta, Original y Bootstrap,
          con su error standart.  \n");
    printf("                                                \n");
    for (fil=0;fil<(m1+1);fil++) printf("Compnte.: %d Original: %f
    SE: %.5f Bootstrap: %f SE: %.5f \n",fil, tdatos_ori[fil],
    se_aux1[fil], tdatos_nag[fil],se_aux2[fil]);
    printf("                                          \n");
    printf("Datos IP Reales y Bootstrap en fila %d valen: %f y
          %f\n",aux1,i_p_ori[aux1],i_p[aux1]);
    printf("                                          \n");
    printf("DEVIANCE's de los modelos Original, Bootstrap e IP:
          %.4f %.4f %.4f\n",dev_ori,dev,dev_err);
    printf("                                        \n");
    printf("El Factor Corrector y su Error valen: FC: %f  ERROR:
          %f\n ",mat[cont1],se_err[1]);
    //printf("                                          \n");
    //printf("bootstrap actual, vale: %f\n ",mat[cont1]);
    printf("                                        \n");
    system("Pause");
    }
  //system("cls");
   gotoxy(10,25);
 if (((cont1==ceil((10*B)/2)+1)||(cont1==1)))
   {//gotoxy(10,30);
    printf("NUEVAS GENERACIONES EN MARCHA. ESPERE,
          POR FAVOR..... ");
    //system("cls");
   }

  cont1++;
   }

  }
while (cont1<(10*B+1));
cont1=0;
cont=0;
vale0=0;
system("cls");
gotoxy(12,6);
printf("          ALGUNOS ERRORES PARCIALES.
                  \n");
```

```
printf("                                    \n");
cont1=0;
cont=0;
vale2=0;
for (cont1=0; cont1<10*B; cont1++)vale2+=fabs(mat[cont1]);
vale2=vale2/(10*B);

for (cont1=0; cont1<10*B; cont1++)
{
 if (mat[cont1]!= 0) vale0+=fabs(mat[cont1]);
  else cont++;
  //system("cls");
if ((cont1==ceil((10*B)/2))||(cont1==0)||
(cont1==ceil((10*B)/2)+4)||(cont1==3)||(cont1==10*B-3)||
(cont1==ceil((10*B)/2)+5)||(cont1==4)||(cont1==10*B-4)||
(cont1==10*B-2)||(cont1==10*B)||(cont1==10*B-1))
{
printf("        El FC en la generacion bootstrap  %d    vale:
              %f\n",cont1,mat[cont1]);
}
//system("Pause");
}
printf("                                    \n");
system("Pause");
cont1=0;
cont=0;
vale0=0;
do
{
 for (cont1=0; cont1<10*B; cont1++)
 aux0=0;
 vale0=mat[cont1];
 for (cont1=0; cont1<10*B; cont1++) {if (mat[cont1]<vale0)
                                    {
                                     aux0=cont1;
                                     vale0=mat[cont1];
                                     };
                                    };
 mat[aux0]=0;
 cont++;
 }
 while (cont<ceil(0.1*B));

for (cont1=0; cont1<10*B; cont1++)
{
 if (mat[cont1]!= 0) vale0+=fabs(mat[cont1]);
  else cont++;
 }
```

```
        vale1=(cont1+1)-cont;
        //printf("El número de valores considerado es: %f\n",vale1);
        //system("pause");
        vale0=vale0/vale1;
        system("cls");
        fin=time(NULL)-3;
        gotoxy(15,10);
        printf("El Factor Corrector vale:          %f\n",vale2);
        printf("                                   \n");
            printf("          El Factor Corrector,corregido, vale:
        %f\n",vale0);
        printf("                                   \n");
        gotoxy(15,15);
        printf("El programa se ha ejecutado en segundos %.1f \n",
        difftime(fin,ini));
        printf("                                   \n");
        printf("                                   \n");
        printf("                                   \n");
        printf("                                   \n");
        printf("                                   \n");
          system("Pause");

      // Final del programa.
        system("cls");
        gotoxy(20,20);
        printf ("Programa realizado por E. Nunnez\n");
        _sleep(1);
        system("cls");
        fclose(f1);
        fclose(f2);
}
```

Bibliografía.

AGRESTI, A. (1996*). An Introduction to Categorical Data Analysis*. Wiley.

ALBERT, A. and ANDERSON, J. A. (1984). On the Existence of Maximun Likelihood Estimate in Logistic Regression Model. *Biometrica.* **71**, 1-10.

ALTMAN, G. (1981). *Practical Statictics for Medical Recharch.* Chapman and Hall.

APPLETON, D. R., (1995). What do we mean by statistical model?. *Statistical in medicine.* **14**, 185-197.

AZZALINI A. (1981). A note on the estimation distribución functional and quantiles by kernel method. *Biometrika, **68**, 326-328.*

B. BOYER, CARL. *Historia de la matemática.* Alianza.

BANEGAS, J. R, VILLAR. F, PÉREZ DE ANDRÉS, C, JIMÉNEZ, R. Estudio epidemiológico de los factores de riesgo cardiovascular en la población española de 35 a 64 años. *Rev San Hig Pub.* **67**, 419-445.

BANEGAS, J. R, RODRÍGUEZ, F, DE LA CRUZ, J. J, DE ANDRÉS, B, DEL REY, J. (1999). Mortalidad relacionada con la hipertensión y la presión arterial en España. *Med Clin.* **112**, 489-494.

BERNSTEIN, I.H. (1988). *Applied Multivariante Analysis.* Springer-Verlag.

BICKEL , P. J. and DOKSUN, K.A. (1987). *Mathematical Statistics*. Holden-Day.

BONAA, K. H., THELLE , D.S., (1991). Association between blood pressure and serum lipids in a population. The Tromso Study. *Circulation.* **83**, 1305-1313.

BONITA, R. (1992). Epidemiology of Stroke. *Lancet.* **339,** 342-344.

BOWMAN, A.W., AZZALINI, A. (1997). Applied Smoothing Techniques for Data Analysis: The Kernel Approach With S-Plus Illustrations. : *Oxford University Press* (pp. 208). ISBN: 0-19-852396-3. OXFORD, UK (UNITED KINGDOM).

BOWMA,N. A., HALL, P. , PRAVAN, T. (1998). Bandwidth selection for the smoothing of distribution funtions. *Biometrica*, **85**,799-808.

BRESLOWN, N.E., DAY, N.E.(1980). *Statistical Methods in Cancer Research.* Vol. 1-The analysis of caee-control studies. International Agency on Cancer, Lyon, France.

BRUNNER, H.R., LARAGH, J.H., BAER, L. (1972). Essential hypertension, renin and aldosterone, heart attack and stroke. *N Engl J Med.* **286,** 441-449.

BUHLER, F.R., TKACHULK, V.A., HAHN, A.W.A., RESINK, T.J. (1991). Low and high density lipoproteins as hormonal regulators of platelet, vascular endothelial and smooth muscle cell interactions: relevance to hypertension. *J Hypertens.* **9**, 172-173.

BYTH, K., MCLACHLAN, G.J. (1988). The biases associated with maximun likelihood methods of estimation of the multivariate logistic risk function. *Commun. Statist. Theory Meth.* **A7**, 877-890.

CAMPBELL, M.J. and MACHIN, D. (1983). *Medical Statistics.* Wiley.

CAO, R., CUEVAS, A., and GONZÁLEZ MANTEIGA, W. (1994), A comparative study of several smoothing methods in density estimation. *Computational. Satistics and Data Analysis.*, **1**, 153-176.

CASTELLI, W.P., ANDERSON, K. (1986). A population at risk. Prevalence of high cholesterol levels in hypertensive patients in the Framingham study. *Am J Med.* **80 Supl 2ª**, 23-32.

CIA GOMEZ, P. (2000). Dislipemias e hipertensión arterial: asociación o coincidencia. *Med Clin.* **115**, 58-59.

COCA, A., GABRIEL, R., DE LA FIGUERA, M., LOPEZ SENDÓN, J.L., FERNÁNDEZ, R., SAGASTAGOITIA, J.D. et al. (1999). The impact of different echocardiographic criteria on prevalence of ventricular hypertrophy in essential hypertension: de Vitae study. *J Hypertens.* **17**, 1471-1480.

Comité de Redacción ad hoc del Grupo de Estudios de Enfermedades Cerebrovasculares de la Sociedad Española de Neurología.(1998). Manejo del Infarto Cerebral en Fase Aguda. *Neurología*; **13 (Supl 3)**: 13-23.

COPAS, J.B.(1983). Regession , prediction and shinkage (whith discussion) *J. R. Statist. Soc .Series* B, **45**, 311-354.

COSÍN, J., VELASCO, J.A., LÓPEZ-SENDON, J.L., DE TERESA, E., DE OYA, M. (1999). Prevención secundaria en España PREVESE 98. ¿Qué ha cambiado?. *Rev Esp Cardiol.* **52 (Supl 4),** 55.

COX, D.R. (1970). *The Analisis of Binary Data.* Metheuen,London.

COX, D.R., HINKLEY, D.V., (1974). *Theoretical Statisctical.* Chapman and Hall

COX, D.R. and SNELL, E.J. (1989). *Analysis of Binary Data.* Chapman and Hall.

CUADRAS, C.M. (1981), *Análisis Multivariante.* Eunibar.

CURB, J.D., PRESSEL, S.L., CYTLER, J.A., et al. (1996). Effect of diuretic-based antihypertension treatment on cardiovascular disease risk in older diabetic patients with isolated systolic hypertension. *JAMA.* **276**, 1886-1892.

DEVIDAS, M., GEORGE, E.O. and ZELTERMAN, D. (1992). Generalized logisted Model for Low-Dose Response Data. *Statistics in Medicin.* **11**, 881-892.

DÍEZ, J. (1994). Current work in the cell biology of left ventricular hypertrophy. *Curr Op Cardiol.* **9**, 512-519.

DORTA, J., PÉREZ, H., BAUTISTA, J. (1986). La hipertensión arterial en la isla de Tenerife. I. Frecuencia. En: Pardel H, Ed La Hipertensión Arterial en España. *Liga Española para la Lucha contra la Hipertensión Arterial,* 155-161.

EFRON, B. (1979). Bootstrap methods: another look of the jacknife. *Ann. Statist.,* **7,** 1-26.

EFRON, B. (1983) "Estimating de Error Rate of a Prediction Rule: Improviment on Cross-Validation." *Journal of the American Statistical Association,* **78,** 316-331.

EFRON, B., TIBSHIRANI R.J., (1993). *An Introduction to the bootstrap.* Chapman & Hall.

EMERSON, J.D. and COLDITTZ, G.A. (1983). Use of Statistical Analysis in the New England. *Journal of Medicine. N. Engl. J. Med.* **309**, 709-713.

FACCHINI, F., IDA CHEN, Y.D., CLINKINGBEARD, C., JEPPESEN, J., REAVEN, G.M. (1992). Insulin resistance and dyslipemia in obese individuals with a family history of hypertension. *Am J Hypertens* . **5**, 694-699

FONTBOUNE, A., ESCHWEGE, E. (1987). Diabetes, hyperglycaemia, hiperinsulinemia and atherosclerosis: epidemiological data. *Diabetes Metab.* **13**, 350-353.

FROHLICH, E.D. (1999). Risk mechanism in hypertensive heart disease. *Hypertension.* **34**, 782-789

FROST, W.H.,(1941) *Papers of Wade Hampton Frost.* The Commonwealth Fund.

FULLER, J.H., MCCARTNEY, P., JARRET, R.J. (1979). Hyperglycaemia and coronary heart disease: the Whithall Study. *J Chron Dis.*, **32**, 721-728.

FUSTER, V., MC GOON, C. (1987). Coartacion de aorta. Cardiology: Fundamentals and Practice. Ed. Bradenburg R, Fuster V, Ginliani E, Mc Goon D. *Years book medical Publishers.* 1438-1444.

GLEIBERMAN, L., (1973). Blood pressure and dietary salt in human population. *Ecology and Food Nutrition.* **2**, 143-156.

GONZALEZ JUANATEY JR, ALEGRÍA E, LOZANO JV, LISTERRI JL, GARCÍA-ACUÑA JM, GONZÁLEZ MAQUEDA I.(200). Impacto de la hipertensión en las enfermedades cardiovasculares en España. Estudio CARDIOTENS 99. *Rev Esp Cardiol* 2000

GONZÁLEZ-VILLALPANDO, C., STERN, M.P., HAFFNER, S.M., GONZÁLEZ-VILLALPANDO, M.E., GASKILL, S., RIVERA MARTINEZ, D. (1999). Prevalence of hypertension in a Mexican population according to the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure. J *Cardiovasc Risk* **6**, 177-181.

GORDON, T., CASTELLI, W.P., HJORTLAND, M.C. (1977) Predicting coronary heart disease in midle-aged and older persons: the Framingham Study. *JAMA.* **238,** 497-499

GREEN, P.J. (1987). Penalized likelihood for general semiparametric regression models. *Internacional Statistical Review.* **55**, 245-259.

HALL, P. (1986). On number of bootstrap simulations requerid to construct a confidencial interval. *Annals of Statistics*, **4**, 1453-1462.

HALL, P. (1992). *The bootstrap and Edgeworth expansión.* Springer-Verlag.

HÄRDLE, W. (1999). *Smothing Tecniques with Implementation in S.* Springer-Verlag.

HAUCK, W.W. and DONNER, A. (1997). Wald's Test Aplications to Hypothesys in logit Analysis. *Jornal of the American Statistical Association.*, **72**, 851-853.

HAYAKAWA, H., RAIJ, L. (1999). Relationship between hypercholesterolaemia, endothelial dysfunction and hypertension. *J Hypertens.* **17**, 611-619.

HIRJI, K.F., TANG, M.L. VOLLESET, S.E., and ELASHOFF, R.M. (1994). Efficient Power Computation for Exact and Mid-P Test for the Common ODDS Ratio in Several 2x2 Tables. *Satatistics in Medicine,* **13,** 1539-1549.

HOERL, A.E., KENNARD, R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

HOLLENBERG, N.K., WILLIANS, G.H. (1995). Abnormal renal function, sodium-volume homeostasis, and renin system behaviour in normal-renin essential

hypertension: The evolution of the non-modular concept, in JH Laragh et al: Hypertension. Pathofhysiology, Diagnosis, and Management, *2d ed. New York Raven Press*, **2**, p 1837.

HOSMER , D., LEMESHOW, S. (1988). *Applied Logistic Regression.* Wiley.

JENICEK, M., CLÉROUX, R.(1993) *Epidemiología. Principios. Técnicas. Aplicaciones.* Masson-Salvat Medicina.

Joint National Committee on Detection, Evaluation and Treatment of High Blood Pressure. The Sixth Report of the Joint National Committee on Detection, Evaluation and Treatment of High Blood Pressure. Arch Intern Med 1997; **157**: 2413-2446

KANNEL, W.B., (1975). Role of blood pressure in cardiovascular disease: The Framingham Study. *Angiology,* **26**, 1.

KANNEL, W.B., DAWLER, T.R., MCGEE, D.L., (1980) Perspectives of Systolic hypertension. *The Framingham Study. Circulation.* **61**, 1179.

KANNEL, W.B., MCGEE, D.L., CASTELLI, W.P. (1984). Latest perspective on cigarette smoking and cardiovascular disease: The Framingham Study. *J Cardiac Rehabil.* **4**, 267-277.

KANNEL, W.B., (1999). Historic perspectives on the relative contributions of diastolic and systolic blood pressure elevation for cardiovascular risk profile. *Am Hearth J. **138***, 205-210.

KAPPENMAN, R.F. (1987). A Nonparametric data base univariate function estimate. *Comput. Statist. Data Anal.* **5**, 1-7.

KATE L.P., BOMAN, H., DAIGER, S.P., (1982) Familial aggregation of coronary heart disease and its relation to known genetic risk factor. *Am J Cardiol.* **50**, 954-53.

KEULEN, E.T., VOORS-PETTE, C., DE BRUIN, T.W. (2001). Familial dyslipidemic hypertension syndrome: familial, combined hyperlipidemia, and the role of abdominal fat mass. *Am J Hypertens.* **14,** 357-363.

KLEINBAUN D.G., KUPPER, L.L., MORGENSTERN, H. (1982). *Epidemilogic Research-Principes and Cuantitative Methodos.* Van Nostrad Reihold Comapany.

KLEINBAUN, D.G. (1993). Una introducción al análisis de regresión logística. *Revisión en Salud Pública.* **3**, 61-105.

LAST, J.M. (1989). *Diccionario de epidemiología.* Salvat Editores.

LEBART, S.L., MORINEAU, A. Y FENELON, A. (1985). *Tratamiento Estadístico de Datos.* Marcombo.

LILIENFIELD, A.M., LILIENFIELD, D.E. (1987). *Fundamentos de epidemiología.* Addison Wesley Iberoamericana.

LÓPEZ GIMENEZ, M.R. (1984). El modelo de regresión logística. Utilización en el campo de la epidemiología. *Cuad. Bioest. Apli. Infor.* **12**, 118-132.

MAC MAHON, B., PUGH, T.F. (1975). *Principios y métodos de epidemiología.* La Prensa Médica Mexicana.

MANLY, B. (1986). *Multivariante statistical method.* Chapman and Hall.

MARRON, J.S. (1989). Common on a data based bandwidth selector. *Comp. Statist. Data Anal.* **8**, 155-170.

MARTIN ANDRÉS, A. y LUNA DEL CASTILLO, J. (1992). *Bioestadística para las ciencias de la salud.* Norma.

MCCULLAGH, P. and NELDER, J.A. (1983). *Generalized linears Models.* Chapman and Hall.
MCFASE-SMITH, W. (1977). Epidemiology of hypertension. *Med Clin North Am.* **61**, 467-486.

MORRISON, D.F. (1976). *Multivariante Statistical Methods.* McGraw-Hill.

MORTON, R.F., HEBEL, J.R. (1987). *Bioestadística y epidemiología.* Interamericana McGraw-Hill.

PARZEN, E. (1962). On estimate of a probability density function and mode. *Annals Mathematical Statictics.* 33/1962, 1065-1076.

PEDRO-BOTET, J. (2002). Hipertensión arterial y dislipemia. *Revista de hipertensión arterial para Atención Primaria.* **25**, 6-25.

PEÑA SÁNCHEZ DE RIVERA, D. (1986). *Estadística: Modelos y métodos (Tomo I).* Alianza.

PREBIGON, D., (1980) Goodness-of-link tests for generalized linear models. *Appl. Statist.* **29**,15-24.

PREBIGON, D.(1981). Logistic Regression diagnostic. *Annals of Staticstis*, **9**, 705-724.

PRIESTLEY, M.B. and CHAO, M.J. (1972). Non-parametric functions fitting. *Journal of the Royal Statistical Society*, **34**, 385-392.

REY CALERO, J. (1989). *Método epidemiológico y salud de la comunidad.* Interamericana McGraw-Hill.

ROSENBLATT, M. (1956). Remarks on some non-parametric estimates of a density function. *Annals of Matematical Statistics*, **27**, 642-669.

ROYSTON, P. (1992). The Use of Cosums and Other Techniques in Modelling continuos Covariates in Logistic Regession. *Statistics in Medicina*, **11**, 1115-1129.

SAAVEDRA, P. (2000). Roc Curves Determination by Nonparametric Methods. *Revista de la Academia Canaria de Ciencias.* Volumen **XII**, Núm. 1-2.

SAN MARTIN, H., MARTIN A.C., CARRASCO, J.L., (1990). *Epidemiología. Teoría Investigación Práctica.* Díaz de Santos.

SÁNCHEZ GARCÍA, M. (1978). Métodos *estadísticos aplicados al Tratamiento de Datos.* CCUCM.

SAVAGE, D.D., DRAYER, J.I., HENRY, W.L., MATHEWS, E.C., WARE, J.H., GRADIN, J.M. et al. (1979). Echocardiography assessment of cardiac anatomy and function in hypertensive patients. *Circulation.* **59**, 623-632.

Servicio del Plan de Salud e Investigación del Servicio Canario de Salud. Encuesta Nutricional de Canarias 1997-1998. Volumen 2: Factores de riesgo cardiovascular. Consejería de Sanidad y Consumo del Gobierno de Canarias. Santa Cruz de Tenerife, 1999. ISBN: 84-89454-22-1.

SHEP Co-operative Research Group. Prevention of Stroke by antihypertensive drug Treatment in older persons with isaolated systolic hypertension. Final results on the systolic hypertension in the elderly program (SHEP). *JAMA* 191; **265**, 3255-3264.

SHOUKRI, M.M. MARTIN, S.W. and MIAN, I.U.H. (1955). Maximum Likelihood Estimates of de kappa Coeficient form Models of Matched Binary Responses. *Statistics in Medicine*, **14**, 83-89.

SILVAPULLE, M.J. (1981). On de existencia of maximun likelihood estimates for de binimial responses models. *J. R. Statist. Soc.* B, **43**,310-313.

SIMONOFF, J.S. (1996). *Smothing Methods in Statistics.* Springer-Verlag.

Sociedades Españolas de Geriatría, Cardiología, Medicina Familiar y Comunitaria, y Liga Española contra la Hipertensión Arterial. ECEHA. (1996 Estudio Cooperativo Español sobre Hipertensión Arterial en el Anciano. Barcelona: Edipharma.

STAESSEN J.A., FAGARD R.,THIJS L., CELIS H., ARABIDZE G.G., BIRKENHAGER W.H., BULPITT C.J., DE LEEUW P.W., DOLLERY C.T., FLETCHER A.E., FORETTE F., LEONETTI G., NACHEV CH., O'BRIEN E.T., ROSENFELD J., RODICIO J.L., TUOMILEHTO J., ZANCHETTI A., for the Systolic Hypertension in Europe (Syst-Eur) Trial Investigators 1997 Randomised double-blind comparison of placebo and active treatment for older patients with isolated systolic hypertension. *The Lancet*, 1997; **350**, 757-764.

STEYERBERG, E.W., EIJKEMANS, M. J.C., HARREL, F.E., HABBEMA J.D.F. (2000). Pronostic modelling with logistic regression analysis: a comparison of selection methods in small data set. *Statistics in medicine.* **19**, 1089-1079.

STEYERBERG, E.W., EIJKEMANS, M.J.C., HABBEMA, J.D. F.(2001). Application of Shrinkage Techniques in Logistic Regression Analysis: A Case Study. *Statistica Neerlandica.* **55,** 76-88.

The Framinghan Study: An epidemiological investigation of cardiovascular diseases. Section 26. US Government Printing Office No. O-414-297. Washinton DC, 1971.

The Pooling Project Research Group.(1978). Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: final report of the Pooling Project. *J. Chron. Dis.* **31**, 201-306.

TIBSHIRANI, R. (1996). Regression Shinkage and selection via the lasso. *J. R. Statist. Soc .Series* B, **58**, 267-288.

TUOMILETHO, J., RASTENYTE, D., BIRKENHAGER, W.H., et al. (1999) Effects of calcium-channel blockade in older patients with diabetes oan systolic hypertension. *N Engl J Med*, **340**, 677-684.

VAN HOUWELIGEN, J.C. & LE CASSIE, S.(1990). Predictive value of statistical models. *Stat Med* , **9,** 1303–1325

WALKER, S.H. and DUNCAN, D.B. (1967) Estimacion of the probability of an event as a function of several indepent variables. *Biometrika*, **54**, 167-169.

WAND, M.P. and JONES, M.C. (1995). *Kernel Smoothing.* Chapman and Hall.

WEBER, K., ANVERSA, P., ARMSTRONG, W. (1992). Remodelling and reparation of the cardiovascular system. *J Am Coll Cardiol.* **2**, 20:3-16.