Trabajo Fin de Máster

Estudio clínico-epidemiológico de cáncer de mama

Nahir Queiro Alberte

Máster en Técnicas Estadísticas

Curso 2023/2024

Propuesta de Trabajo Fin de Máster

Título en galego: Estudo clínico-epidemiolóxico de cancro de mama
Título en español: Estudio clínico-epidemiológico de cáncer de mama
English title: Clinical-epidemiologic study of breast cancer
Modalidad: Modalidad B
Autor/a: Nahir Queiro Alberte, Universidad de Santiago de Compostela
Director/a: Paula Saavedra Nieves, Universidad de Santiago de Compostela
Tutor/a: Manuela Gago Domínguez, Fundación Pública Gallega de Medicina Genómica
Breve resumen del trabajo: En este trabajo se realiza un análisis estadístico del cáncer de mama sobre una base de datos aportada por la Fundación Gallega de Medicina Genómica. Después de realizar un análisis descriptivo exhaustivo, se emplea el modelo de regresión logística para determinar qué variables suponen un factor de riesgo o de protección. Finalmente, estudiamos la utilidad de los modelos ajustados para la predicción de nuevos casos.
Recomendaciones:
Otras observaciones:

Doña Paula Saavedra Nieves, profesora de la Universidad de Santiago de Compostela, informa que el Trabajo Fin de Máster titulado

Estudio clínico-epidemiológico de cáncer de mama

fue realizado bajo su dirección por doña Nahir Queiro Alberte para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 1 de junio de 2024.

El/la director/a:

SAAVEDRA NIEVES PAULA - 44847364D

Firmado digitalmente por SAAVEDRA NIEVES PAULA - 44847364D
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-44847364D, givenName=PAULA, sn=SAAVEDRA NIEVES, cn=SAAVEDRA NIEVES PAULA - 44847364D
Fecha: 2024.06.02 10:32:27 +02'00'

Doña Paula Saavedra Nieves

El/la autor/a:

QUEIRO ALBERTE NAHIR - 53796042P

Firmado digitalmente por QUEIRO ALBERTE NAHIR - 53796042P
Fecha: 2024.06.02 18:44:09 +02'00'

Doña Nahir Queiro Alberte

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Agradezco a la directora del trabajo, Paula Saavedra Nieves, su atención y apoyo durante la elaboración, así como todas sus propuestas para el análisis estadístico de la muestra, sin las cuales no habría sido posible realizar esta memoria. Al Consorcio Gallego de Cáncer de Mama, BREOGAN, en especial a la Fundación Gallega de Medicina Genómica, por facilitarme la base de datos de cáncer de mama utilizada. Por último, a mi familia y amigos por sus ánimos y apoyo durante todo el proceso.

Índice general

Resumen

Resumen en español

La epidemiología estudia la distribución y los fenómenos o determinantes relacionados con la salud, con el fin de conocer la causa o causas de una enfermedad y mejorar su diagnóstico y tratamiento. En este trabajo realizamos un estudio estadístico sobre el cáncer de mama, revisando en primera instancia estudios previos sobre esta enfermedad, ampliamente estudiada. Posteriormente, a partir de una base de datos de cáncer de mama proporcionada por la Fundación Gallega de Medicina Genómica, hemos realizado un análisis exploratorio y se han ajustado distintos modelos de regresión logística para identificar qué variables suponen un factor de riesgo o de protección del cáncer de mama. La base de datos está compuesta por 672 observaciones de casos (300) y controles (372), que cuenta con un total de 36 variables relacionadas con la edad, la menarquia, la menopausia, con factores genéticos, con hábitos de vida, con características antropométricas y con la concentración de algunos tipos de células en sangre. Finalmente, hemos estudiado la capacidad diagnóstica de los modelos de regresión logística empleando curvas ROC. Para todo ello se ha hecho uso del programa RStudio.

English abstract

Epidemiology studies the distribution and phenomena or determinants related to health, with the aim of understanding the causes of a diseases and improving their diagnosis and treatment. In this work we carried out a statistical study on breast cancer, first reviewing previous studies on this disease, which has been widely studied. Subsequently, using a breast cancer database provided by the Galician Breast Cancer Consortium, BREOGAN, we performed an exploratory analysis and fitted different logistic regression models to identify which variables are risk or protective factors for breast cancer. The database is composed of 672 observations of cases (300) and controls (372), which has a total of 36 variables related to age, menarche, menopause, genetic factors, lifestyle habits, anthropometric characteristics and the concentration of some cell types of blood. Finally, we have studied the diagnostic

capacity of logistic regression models using ROC curves. RStudio software was used for this purpose.

Preliminares

El cáncer de mama es una de las enfermedades con mayor incidencia y mortandad en todo el mundo. En el año 2020 superó al de pulmón como el cáncer con mayor número de diagnósticos (Bray et al., 2018), con una estimación de 2.3 millones de casos nuevos con respecto al año anterior. De la totalidad de diagnósticos de cáncer durante ese año, el cáncer de mama recoge al 11.7 %, y es el quinto con mayor índice de mortandad (el 6.9 %) (Sung et al., 2021). El riesgo acumulado de padecer esta enfermedad es del 5.2 % a la edad de 74 años, con un 1.5 % de probabilidades de causar la muerte. En Sung et al. (2021), encontramos la Figura 1, donde mediante varios gráficos de sectores se muestra la incidencia y la mortandad de los cánceres más comunes por sexo. Efectivamente, entre las mujeres, el cáncer de mama es el más diagnosticado y el que más muertes ha causado en el año 2020.

La incidencia de los diagnósticos es mucho mayor en países que pertenecen a regiones más desarrolladas que aquellos en vías de desarrollo, tal y como podemos observar en el gráfico de barras de la Figura 2, disponible en Sung et al., (2021). En ella se representa la incidencia y la mortandad de este cáncer por regiones. Este fenómeno puede deberse a una mayor prevalencia de factores de riesgo hormonales, reproductivos y determinados hábitos de vida que ocurren en mayor medida en países enriquecidos: edad temprana de la menarquia, retraso en la edad de la menopausia, retraso de la edad al primer hijo, menor número de hijos, menor tiempo dando el pecho, uso de anticonceptivos orales y terapias hormonales para la menopausia, consumo de alcohol, sobrepeso y obesidad, sedentarismo... Debemos tener en cuenta, además, que en estos países existen programas de control y prevención del cáncer de mama, lo que aumenta el número de diagnósticos registrados.

Ante esta sobrecogedora realidad, el cáncer de mama ha sido, y es, objeto de numerosos estudios encaminados a descubrir sus causas para posibilitar el desarrollo de diagnósticos precoces y tratamientos efectivos. Gago-Domínguez et al. (2016) concluyeron mediante un estudio de casos y controles, que el aumento en el consumo de alcohol aumenta a su vez las probabilidades de padecer cáncer de mama en 3 de los 4 subtipos de cáncer analizados. Esta asociación estaba presente además en los diferentes grados de cáncer de mama, del I al III, indicadores de la gravedad de la enfermedad. En Ali et al. (2014), se lleva a cabo un meta-análisis de un total de 11 estudios sobre el consumo de alcohol y el cáncer de

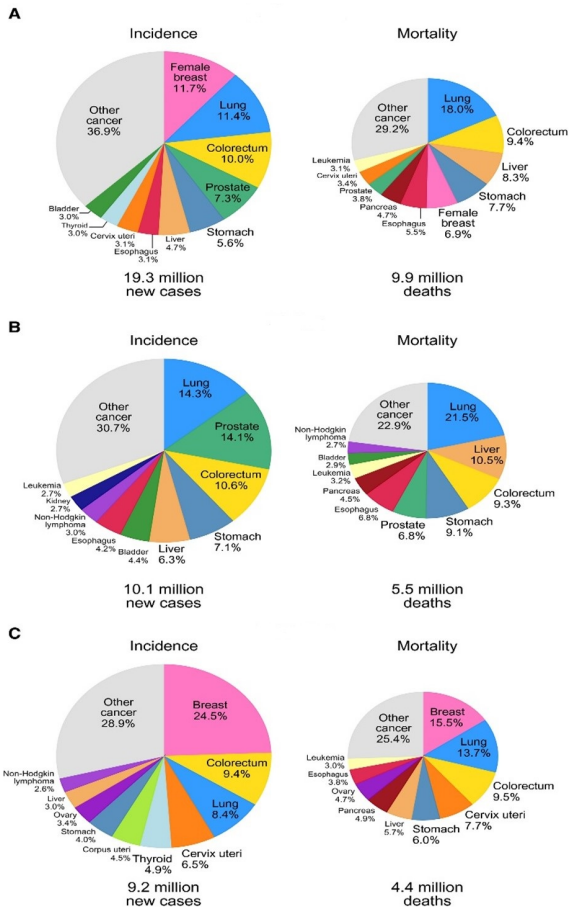


Figura 1: Distribución de los casos y muertes más comunes en el año 2020 (en inglés) para (A) ambos sexos, (B) hombres y (C) mujeres. Fuente: Sung et al. (2021).

mama. Apuntan que este hábito no solo aumenta las probabilidades de padecer dicha enfermedad, sino que una vez diagnosticado, la ingesta moderada de alcohol aumenta las probabilidades de mortandad con respecto a un consumo nulo. Chen et al. (2011) también encontraron relaciones significativas entre el cáncer de mama y el consumo de alcohol incluso en cantidades bajas (de 3 a 6 bebidas por semana). El riesgo aumenta considerablemente cuando los niveles de consumo se sitúan alrededor de 2 bebidas al día, comparado con mujeres que no consumen alcohol. Prácticas como el "bringe drinking" o "consumo por atracón" (consumir en un periodo corto de tiempo 4 o más bebidas alcohólicas), produce los mismos efectos en cuanto al riesgo de desarrollar esta enfermedad en comparación con un consumo más frecuente, aunque moderado.

Distintas variables antropométricas también han sido estudiadas desde la epidemiología del cáncer de mama. Reeves et al. (1996) y Cecchini et al. (2012) identificaron que un mayor Índice de Masa Corporal (IMC) se asocia con un peor pronóstico de la enfermedad y con mayores probabilidades de

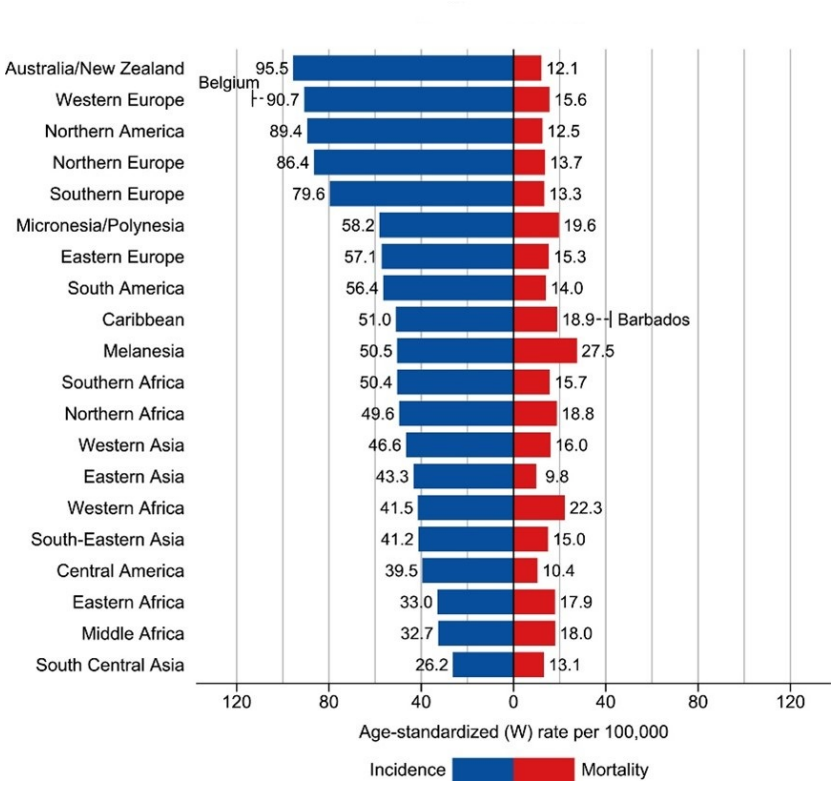


Figura 2: Incidencia y mortandad del cáncer de mama en 2020 por regiones (en inglés), estandarizado por la edad. Fuente : Sung et al. (2021).

recurrencia y muerte. Destacan que la obesidad puede elevar el crecimiento tumoral debido a la influencia de determinadas hormonas, y que, de cara a la prevención, un peso alto dificulta o retrasa la detección de bultos en la mama al ser más difíciles de encontrar mediante la palpación. Suzuki et al. (2009) evidencian que, en esta asociación entre el cáncer de mama y el peso, el estatus menopáusico tiene implicaciones notables. En su meta-análisis de 22 casos-controles y 9 estudios de cohortes, concluyeron que el riesgo de padecer cáncer ER+/PR+ (receptores de estrógenos y progesterona positivos) era un 10 % más bajo para aquellas mujeres en la premenopausia y con un peso mayor al de referencia. Sin embargo, el riesgo aumentaba hasta un 33 % para el caso de mujeres con menopausia por cada 5 unidades incrementadas en su IMC. La causa puede encontrarse en la relación que la obesidad tiene con el estado menopáusico, pues influye de formas diferentes en los estrógenos antes y después de esta. Concuerdan con estos hallazgos Picon-Ruiz et al. (2017) y Trentham-Dietz et al. (1997). Sin embargo, Kawai et al., (2014) y Cecchini et al. (2012) sí encuentran en sus análisis una asociación positiva entre un mayor IMC y el riesgo de cáncer de mama en la premenopausia, especialmente para mujeres que presentan un riesgo previo y cuyo IMC supera los 10 Kg/m^2. Resaltan que esta relación positiva entre el IMC y el riesgo de cáncer de mama en mujeres premenopáusicas ocurre especialmente en edades

mayores a los 35 años, por lo tanto, la relación inversa entre el IMC y el riesgo de cáncer de mama observada por Suzuki et al. (2009), puede darse por la influencia de la edad.

No solo el peso, sino la ganancia de éste, se asocian con un incremento en las probabilidades de padecer cáncer de mama postmenopausia (Trentham-Dietz et al., 1997). Ganancias de al menos 10 Kg desde los 35 años en adelante, aumentan en un 27 % el riesgo de cáncer, comparado con ganancias más bajas. Sin embargo, la pérdida progresiva de peso disminuye estos valores por cada Kg perdido.

Trentham-Dietz et al., (1997) también encontraron asociaciones con la estatura. Esta variable ha sido ampliamente estudiada en relación a muchos tipos de cáncer en Kabat et al., (2012). En su estudio, la altura presenta una relación positiva con el riesgo de cáncer de mama independientemente del estado menopáusico, aumentando en un 36 % en las mujeres más altas de la muestra. Zhang et al. (2015) encontraron evidencias de que cada 10 cm de altura, el riesgo relativo de sufrir esta enfermedad aumenta en un 17 %.

La edad a la menarquia ha mostrado asociaciones positivas y negativas con el riesgo de cáncer de mama. Mujeres con menarquias anteriores a los 13 años tienen un riesgo mayor de padecer esta enfermedad, mientras que las más tardías actúan como factor de protección (Ritte et al., 2012). Los hábitos de vida y de alimentación propios de la sociedad moderna han disminuido la edad de la primera regla, evento que ha favorecido también a un incremento en la altura media de la población (Zhang et al., 2015).

Otro efecto de los cambios en las sociedades modernas es el retraso en la maternidad, factor que contribuye al aumento del número de casos de cáncer de mama a edades tempranas (Cruz et al., 2013). Se ha demostrado que después de la maternidad, existe un riesgo más elevado de padecer cáncer de mama hasta los 10 años siguientes a la fecha de dar a luz, en comparación a mujeres que no han tenido hijos. Sin embargo, una vez pasado este periodo, el riesgo se reduce considerablemente en comparación a mujeres nulíparas. Y no solo la maternidad, un periodo de lactancia igual o mayor a 7 meses disminuye el riesgo (Redondo et al., 2012). Sin embargo, los resultados en cuanto al riesgo atribuible a estos factores cambian considerablemente dependiendo del tipo de cáncer.

También Phipps et al. (2011), Palmer et al. (2011) y Redondo et al. (2012) encontraron asociaciones entre factores reproductivos, menstruales y lactantes con determinados subtipos de cáncer de mama. No haber tenido hijos se asocia con un decrecimiento del riesgo de padecer cáncer de mama triple-negativo (ausencia de receptores de estrógenos y progesterona, así como de la proteína HER2), pero incrementa el riesgo de ER+/PR+, especialmente a partir de los 45 años. Entre las mujeres que sí han dado a luz, el número de nacimientos correlaciona de forma positiva con el riesgo de cáncer triple-negativo, pero es inversamente proporcional al ER+ en la menopausia. La edad a la menarquia y la menopausia han mostrado una modesta asociación con el riesgo de cáncer ER+. La lactancia, sin embargo, supone un factor de protección en estos estudios. Mujeres con muchos partos y lactancias largas, muestran una

reducción en el riesgo de PR+/ER+, aunque tener hijos a edades tardías se asocia con un incremento del riesgo de cáncer para todos los subtipos.

El uso de anticonceptivos orales también parece tener implicaciones en el riesgo de padecer cáncer de mama. Dolle et al. (2009) encontraron que estos aumentan las probabilidades de desarrollar el subtipo triple-negativo en mujeres menores de 40 años que habían usado anticonceptivos orales al menos durante 6 años. Este riesgo aumenta aún más para mujeres que iniciaron su consumo antes de los 18 años, comparado con aquellas que nunca los habían utilizado. Igualmente, Ma et al. (2010) encontraron asociaciones positivas también para mujeres con edades comprendidas entre los 45 y los 64 años que habían empezado a usarlas por primera vez antes de los 18 años.

La historia familiar es también un factor importante, y no sólo en el cáncer de mama, sino en otros muchos. Jiang et al. (2012) y Pharoah et al. (1997) ponen de manifiesto que las mujeres con antecedentes familiares de cáncer de mama tienen más probabilidades de padecer cáncer del subtipo ER-/PR- (receptores negativos de estrógenos y progesterona) antes de los 50 años.

Investigaciones más recientes se han interesado también en descubrir las implicaciones de células como los neutrófilos en esta enfermedad. Estas células son un tipo de glóbulo blanco producido por el sistema inmunitario, destinadas a combatir infecciones como las provocadas por bacterias y virus. En estos estudios se ha encontrado que una cantidad alta de neutrófilos en sangre se asocia a una menor probabilidad de supervivencia y superación de la enfermedad. La ratio neutrófilos/linfocitos se ha estudiado también como un biomarcador de la gravedad del cáncer, su pronóstico y lesiones metastáticas. Iwase et al. (2016), Faria et al. (2016) y Ethier et al. (2017) encontraron que esta ratio, conocida como NLR (*Neutrophil to Lymphocyte Ratio*) aumenta a medida que avanza la enfermedad, especialmente en el subtipo triple-negativo. Un NLR alto se asocia a una menor esperanza de vida y menores probabilidades de recuperación en pacientes con cáncer de mama, especialmente en los subtipos ER-negativo y HER2-negativo. Gago-Domínguez et al. (2020), llegan a conclusiones muy similares en su estudio de casos-controles hecho en Galicia, encontrando asociaciones positivas entre un elevado NLR y un incremento en el riesgo de cáncer de mama, especialmente de los subtipos Luminal A (ER/PR positivos y HER2 negativo) y HER2 negativo. Añaden que esta asociación es más pronunciada en mujeres con altos niveles de colesterol y bajos en H_2O_2 (agua oxigenada).

Uribe-Querol y Rosales (2015) explican que los neutrófilos son los leucocitos más comunes en la sangre, y los primeros en la línea de defensa contra inflamaciones e infecciones. Por esto, su presencia en la sangre cuando se dan enfermedades como el cáncer, es normal. Sin embargo, parece ser que los neutrófilos pueden ser causantes de la infiltración de tumores, esto es, durante el proceso de lucha contra las células tumorales, pueden activar funciones que favorecen el desarrollo y metástasis del cáncer. Faria et al. (2016) añaden que estas células pueden secretar el conocido como *factor de crecimiento endotelial vascular* (VEGF), una sustancia que estimula la formación de nuevos vasos sanguíneos en el tumor y

promueve su desarrollo. Este fenómeno, conocido como angiogénesis tumoral, contribuye, por lo tanto, a la progresión de la enfermedad y explica la correlación negativa entre un alto recuento de neutrófilos y la supervivencia de las pacientes. Un NLR alto retrasa, además, la respuesta a la quimioterapia.

En definitiva, el cáncer de mama es una enfermedad compleja. Los factores de riesgo que pueden estar relacionados con ella son numerosos, y hasta la fecha, no se ha encontrado una causa o causas específicas para su aparición y desarrollo. En Galicia, el Consorcio Gallego de Cáncer de Mama BREOGAN (*BREast Oncology Galician Network*, https://proyectobreogan.es/), trabaja en la investigación de esta patología desde el año 2010. Está formado por profesionales y pacientes de los complejos hospitalarios de Vigo y Santiago de Compostela y por la Fundación Pública Gallega de Medicina Genómica. Sus objetivos principales son determinar el riesgo individual de padecer cáncer de mama, atendiendo a factores tanto genéticos como de estilos de vida mediante estudios de tipo caso-control. Para la elaboración de este Trabajo de Fin de Máster, la Fundación Pública Gallega de Medicina Genómica nos ha facilitado una muestra de 672 mujeres extraídas de la misma base poblacional, de las cuales 300 son casos y 372 son controles. La base de datos resultante cuenta con variables como la edad al diagnóstico o a la entrevista dependiendo de la condición de caso o control, edad a la menarquia, estado menopáusico, edad a la menopausia, maternidad, número de hijos, edad al primer hijo, lactancia, IMC, el uso de anticonceptivos orales y otras muchas más en las que profundizaremos más adelante. Tras realizar un análisis exploratorio exhaustivo, emplearemos el modelo de regresión logística con el objetivo de identificar qué variables suponen un riesgo de cara al desarrollo de esta enfermedad, y así predecir las probabilidades individuales de sufrir cáncer de mama en base a la información aportada por los datos disponibles. La capacidad diagnóstica de los modelos logísticos ajustados sera evaluada empleando curvas ROC (*Receiver Operating Characteristic curve*).

Esta memoria se organiza cómo sigue. En el Capítulo 1 se revisa el modelo logístico y su importancia en el estudio epidemiológico, centrándonos en conceptos como la incidencia, la prevalencia, el riesgo relativo y la odds ratio. Revisaremos los fundamentos principales de este tipo de modelos y los métodos utilizados para la estimación de sus parámetros, así como de otros aspectos relacionados con la selección de variables y la evaluación del modelo como test diagnóstico. El Capítulo 2 presenta los resultados del análisis exploratorio de la base de datos facilitada por la Fundación Pública Gallega de Medicina Genómica, con gráficos y cuadros que muestran información sobre la distribución y las medidas de posición y dispersión de las variables. Se muestran además, los modelos de regresión logística ajustados para el análisis de las variables determinantes en el desarrollo del cáncer de mama y se analiza la capacidad de predicción de casos nuevos de los modelos. El Capítulo 3 contiene las conclusiones extraídas del trabajo sobre la investigación del cáncer de mama y sus posibles causas (factores de riesgo y protección), las dificultades encontradas durante el análisis de la base de datos, y reflexiones sobre los modelos de regresión ajustados. Por último, encontramos material suplementario

en los apéndices. En el Apéndice A, se muestran representaciones gráficas adicionales no contenidas en el Capítulo 2. En el Apéndice B, encontraremos el recuento de valores perdidos de la base de datos. En el Apéndice C, se muestran cuadros con los ajustes de modelos de regresión no válidos para el análisis pero que sí se han tenido en cuenta para el ajuste de los modelos propuestos. En el Apéndice D, aparecen las pruebas de linealidad y las distancias de Cook aplicadas a los modelos escogidos. Por último, en el Apéndice E, se añade el código de RStudio utilizado para los análisis.

Capítulo 1

La regresión logística y su aplicación en la epidemiología

En el campo de la investigación médica, la epidemiología estudia la distribución y los fenómenos o determinantes relacionados con la salud, con especial interés en su frecuencia, causas, factores de riesgo y modos de trasmisión (Bonita et al., 2008). Su objetivo principal es, por lo tanto, valorar y medir eventos relacionados con la salud y las enfermedades, haciendo uso de métodos estadísticos con los que valorar una situación relativa a la salud de un grupo o colectivo poblacional (Mirón y Alonso, 2008). Su trabajo es fundamental de cara al desarrollo de métodos preventivos y la mejora de medidas sanitarias destinadas a aumentar los niveles de salud poblacional.

La frecuencia que presenta una enfermedad es fundamental en la toma de decisiones profesionales. Las dos medidas más utilizadas para contabilizarla o cuantificarla son la incidencia y la prevalencia. La **incidencia** mide la frecuencia o aparición de una enfermedad en un periodo de tiempo, mientras que la **prevalencia** establece la proporción de sujetos que tienen una enfermedad en un momento determinado, sin tener en cuenta el paso del tiempo (Mirón y Alonso, 2008).

La incidencia solo puede ser medida en estudios prospectivos, pues necesitamos partir de un número de sujetos susceptibles de desarrollar la enfermedad en estudio, y observar hacia el futuro cuantos de ellos finalmente enferman (Cerda et al., 2013 y Aguilar et al., 2017). Por otro lado, el Instituto Nacional del Cáncer (NCI, `https://www.cancer.gov/espanol`) define la prevalencia como una "medida del número total de personas en un grupo específico que tienen (o tuvieron) cierta enfermedad, afección o factor de riesgo en un momento específico o durante un periodo determinado". Se trata de un indicador puntual que no incluye el factor tiempo y que establece la carga de una enfermedad en una determinada población (Mirón y Alonso, 2008). Es la medida idónea para estudios retrospectivos y transversales,

en los que no es posible introducir la variable tiempo para determinar la proporción de sujetos que padecerán la enfermedad en un futuro. En un estudio retrospectivo, la muestra es escogida a partir de un evento de interés, en este caso, sujetos que tienen una determinada enfermedad, y se indaga en su pasado la presencia de una determinada exposición. En un estudio transversal, se selecciona una muestra sin conocer *a priori* la condición de salud de los sujetos y su exposición al factor de riesgo, características que se determinan *a posteriori* y se miden de forma simultánea. Al carecer de un sentido de observación prospectivo, estos tipos de estudios no permiten el cálculo de tasas de incidencia (Cerda et al., 2013), por consiguiente, es más correcto desde un punto de vista metodológico utilizar la prevalencia. Podemos ver en el Cuadro 1.1 elaborado por Mirón y Alonso (2008) las diferencias más notables entre ambos términos.

Incidencia	**Prevalencia**
Indica la probabilidad de desarrollar la enfermedad	Indica la probabilidad de que ya se padezca la enfermedad
Para su cálculo debemos colocar en el numerador sólo los casos nuevos	Para su cálculo, colocamos en el numerador todos los casos, nuevos y viejos
Es necesario el seguimiento de los individuos	No se necesita un seguimiento
Su valor no depende de la duración de la enfermedad	La duración de la enfermedad sí influye en su valor
Valora y cuantifica enfermedades agudas	Valora y cuantifica enfermedades crónicas
Utilizada para investigar y establecer relaciones causales	Utilizada para valorar la carga o coste de una enfermedad crónica

Cuadro 1.1: Diferencias entre incidencia y prevalencia. Fuente: Mirón y Alonso (2008).

Para profundizar en conceptos relacionados con la medición de la incidencia y la prevalencia, este capítulo se organiza cómo sigue. En la Sección 1.1, se presenta el concepto de odds ratio, clave en regresión logística, así como sus diferencias con el riesgo relativo. En la Sección 1.2, se justifica el uso del modelo de regresión logística cuando la variable respuesta es dicotómica, como es nuestro caso. La Sección 1.3 introduce conceptos fundamentales del modelo logístico, su relación con la odds ratio

y la *función logit*. Además, presenta los métodos utilizados para la estimación de los parámetros, la selección de variables y el análisis de la bondad de ajuste. Finalmente, se revisan las curvas ROC como herramientas para evaluar la capacidad diagnóstica de los modelos ajustados sobre de la base de datos aportada por la Fundación Gallega de Medicina Genómica.

1.1. Riesgo relativo y odds ratio

El *riesgo relativo* (RR) y la *odds ratio* (OR) son medidas de efecto utilizadas para conocer la magnitud cuantitativa de la fuerza de asociación entre una variable independiente y otra dependiente. Específicamente, la primera de ellas suele referirse a una exposición o factor de riesgo, y la otra una enfermedad o efecto (Mirón y Alonso, 2008; Cerda et al., 2013 y Szumilas, 2010).

El RR estima la frecuencia de una enfermedad en un grupo de individuos expuestos en relación a un grupo de no expuestos. Es decir, estima la magnitud del efecto (probabilidad de que se desarrolle la enfermedad en cuestión) en el grupo de expuestos a un factor de riesgo en relación al grupo de no expuestos (Mirón y Alonso, 2008). Esta medida se utiliza comúnmente para medir la incidencia en estudios cuyo objetivo es determinar la magnitud de la asociación exposición-enfermedad en términos relativos, como estudios prospectivos de cohortes o ensayos clínicos (Mirón y Alonso, 2008; Cerda et al., 2013). Sin embargo, en estudios donde la población seleccionada ya ha desarrollado la enfermedad, o se desconoce *a priori* esta condición, no es posible utilizar el RR como medida de asociación. Como hemos visto, en estudios retrospectivos y transversales, no sabemos o no se tiene en cuenta el momento en el que aparece la enfermedad, por lo que no es posible medir la incidencia utilizando el RR en este tipo de escenarios. Aquí, es necesaria la búsqueda de medidas alternativas como la OR.

Cerda et al. (2013, p.1330) definen la odds como "un cociente entre el número de eventos y el número de no eventos". En el caso que nos ocupa, se corresponde con el cociente entre el número de enfermos y no enfermos. La OR es una medida que va más allá y añade una variable o factor de exposición, definiéndose entonces como la razón entre la odds cuando ha habido una exposición, frente a la odds cuando no ha habido dicha exposición (Mirón y Alonso, 2008). En epidemiología se utiliza para comparar la posibilidad de enfermar después de la exposición a un factor de riesgo versus la posibilidad de enfermar sin la exposición. Es muy utilizada en estudios de tipo caso-control dado el carácter retrospectivo que presentan este tipo de investigaciones (Aguilar et al., 2017 y Szumilas, 2010).

Podemos observar en el Cuadro 1.2 cómo se relacionan exposición y enfermedad en una tabla de contingencia cuando éste es nuestro evento de interés, donde a, b, c y d, denotan las frecuencias observadas. Concretamente, la cantidad de enfermos expuestos a un factor de riesgo (a), enfermos no expuestos (c), no enfermos expuestos (b), y no enfermos no expuestos (d). En el Cuadro 1.3 se muestran

las fórmulas de ambas medidas de asociación en base a las frecuencias establecidas en el Cuadro 1.2.

Evento de interés: enfermedad

Exposición	**Enfermos**	**No enfermos**	**Total**
Si	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	a + b + c + d

Cuadro 1.2: Tabla epidemiológica. Relación entre exposición y enfermedad.

Medidas de asociación

RR	(a/a+b) / (c/c+d)
OR	(a/b) / (c/d) o (a · d) / (c · b)

Cuadro 1.3: Cálculo del RR y el OR a partir de los términos plantedos en el Cuadro 1.2

A pesar de la diferencia que presentan ambas medidas en base al cálculo de la incidencia, muchos estudios prospectivos utilizan también la OR como medida de asociación. Esto se debe a que muchos de ellos usan la regresión logística como modelo de regresión, que utiliza la OR como medida de efecto, gracias a su capacidad para minimizar lo que se conoce como "sesgo de confusión", un fenómeno que ocurre cuando una variable extraña, externa al estudio, interfiere en la dirección de la asociación entre la exposición estudiada y la enfermedad, dificultando el análisis de relaciones de causalidad (Cerda et al., 2013; y Bonita et al., 2008).

La OR es también muy utilizada como medida de la magnitud de una intervención o tratamiento, puesto que realiza una interpretación contrafactual, es decir, el número de sujetos considerados no interfiere en los resultados, algo que sí sucede con el RR (Aedo et al., 2010).

Las dos medidas, OR y RR, pueden ser considerados equivalentes o muy próximos cuando la

enfermedad o suceso estudiado es muy raro y ocurre en menos del 10 % de los sujetos (Aedo et al., 2010; Cerda et al., 2013; Aguilar et al., 2017 y Mirón y Alonso, 2008). Sin embargo, vemos claramente en la Figura 1.1 extraída de Cerda et al., (2013), cómo a medida que la frecuencia aumenta, la OR refleja con mayor fuerza la magnitud de la asociación entre exposición y enfermedad (Aguilar et al., 2017 y Mirón y Alonso, 2008).

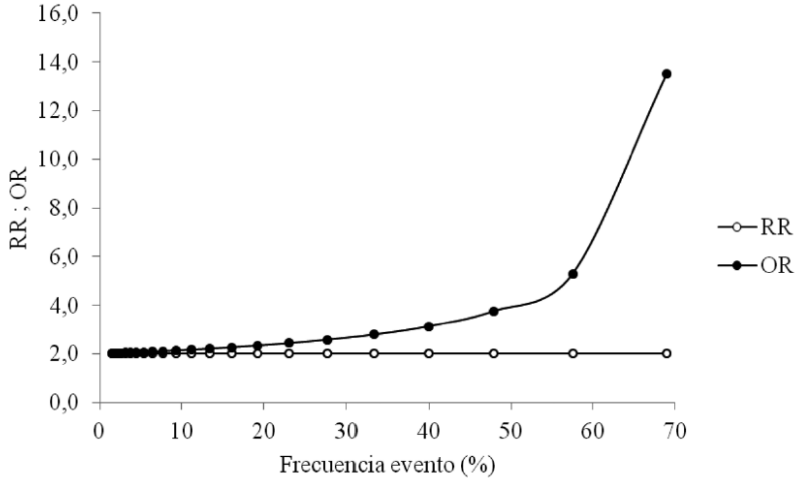


Figura 1.1: Comparación del RR y el OR según la frecuencia del evento de interés. A medida que aumenta la frecuencia, el OR difiere y magnifica más el tamaño del efecto que el RR. Fuente: Cerda et al. (2013).

En este ejemplo, Cerda et al. (2013) enfrentan la frecuencia del evento estudiado (eje X), en este caso la incidencia de una enfermedad, frente a los valores adoptados por las dos medidas de efecto estudiadas (eje Y). Utilizan para ello, una muestra de 100 casos y 100 controles, siendo la proporción de sujetos enfermos y expuestos el doble de la observada en el grupo de no expuestos (RR = 2). Al aumentar la frecuencia de la enfermedad en la muestra, el OR magnifica el tamaño del efecto, adquiriendo valores mucho más altos que el RR.

1.2. Justificación del modelo logístico

Los modelos de regresión son muy utilizados en epidemilogía para analizar la relación existente entre una variable respuesta, en este caso una enfermedad, y una o varias variables explicativas, correspondientes a los factores de exposición. En el campo de la medicina es habitual encontrarnos con variables respuesta de naturaleza dicotómica, esto es, que solo toman dos valores posibles (por ejemplo, ausencia o presencia de una enfermedad). Bajo este escenario, el modelo logístico es el más recomendable para el análisis de este tipo de datos (Hosmer et al., 2013).

Formalmente, estos modelos describen las relaciones existentes entre una variable aleatoria respuesta Y binaria y k variables aleatorias independientes agrupadas en el vector $X = (X_1, X_2, X_3, ..., X_k)$. La variable respuesta Y sólo toma los valores 0 y 1, fracaso y éxito respectivamente, que se corresponderían con el incumplimiento o no de una determinada condición. En el caso que nos ocupa, 0 se corresponde con no sufrir cáncer de mama, y 1 con padecerlo. Dada esta condición, la distribución de la variable Y es Bernoulli y, por tanto, su media es igual a la probabilidad de éxito (en el caso que nos ocupa, probabilidad de sufrir cáncer de mama). Es decir,

$$E(Y) = 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0) = P(Y = 1).$$

Es conocido que la simplicidad de los modelos de regresión lineal puede ser muy útil en el análisis de datos, sin embargo, cuando la variable respuesta Y es dicotómica y se asume un modelo lineal,

$$Y_i = x_i^{'}\widetilde{\beta} + \varepsilon_i \qquad \text{para} \qquad i \in \{1, \ldots, n\},$$

donde Y_i es el valor de la variable respuesta para el individuo i-ésimo, $x_i^{'}$ denota un vector que contiene valores de las k variables explicativas, $\widetilde{\beta}$ es el vector de coeficientes del modelo lineal y ε_i representa el error, que sigue una distribución N(0, σ^2). De este modo, $E(Y_i|X = x_i) = x_i\beta$, por lo que podrían predecirse valores fuera del soporte [0,1] para la probabilidad de sufrir cáncer de mama, lo cual no tiene sentido en este caso. La media condicional de Y debe ser menor o igual a 1 y mayor o igual a 0, esto es, $0 \leq E(Y_i|X = x_i) \leq 1$. Así, los cambios producidos en $E(Y|x)$ por cada unidad de x, serán progresivamente más pequeños a medida que esta media condicionada se acerca a uno o a cero (Hosmer et al., 2013).

En la Figura 1.2 hemos ajustado un modelo de regresión simple con la variable explicativa *ratio neutrófilos/linfocitos* o *NLR* (eje X), presente en la base de datos aportada por la Fundación Gallega de Medicina Genómica, frente a la variable indicadora de cáncer de mama (eje Y). Observamos que la línea ajustaba predice valores por encima y por debajo del soporte [0,1], algo inadmisible en este contexto.

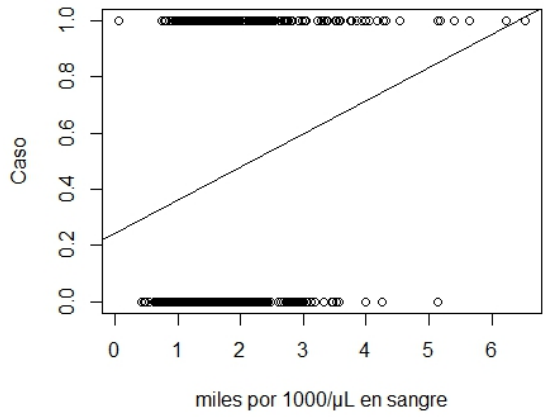


Figura 1.2: Diagrama de dispersión de la variable indicadora de cáncer de mama frente a la ratio neutrófilos/linfocitos, donde se muestra el modelo de regresión simple ajustado.

También podríamos encontrar problemas con la varianza si ajustamos un modelo lineal cuando Y es binaria. Bajo la hipótesis de homocedasticidad, la varianza de las observaciones se supone constante en el modelo lineal, $Var(Y_i|X = x_i) = \sigma^2$. Sin embargo, cuando nuestra respuesta sólo admite dos posibles valores, la varianza: $Var(Y_i|X = x_i) = P(Y_i = 1|X = x_i)[1 - P(Y_i = 1|X = x_i)]$, dando como resultado un modelo heterocedástico totalmente incompatible con un modelo lineal.

Finalmente, la hipótesis de normalidad de los errores propia del modelo lineal tampoco se cumple, pues Y no sigue una distribución normal, sino que es Bernoulli tal y como hemos comentado previamente.

El quebrantamiento de estas tres hipótesis anula la posibilidad de utilizar la regresión lineal como modelo predictivo cuando la variable respuesta es binaria. La mejor alternativa es la regresión logística, que se fundamenta en la OR para el cálculo de probabilidades (Hosmer et al., 2013). Por estos motivos, será el modelo que usaremos en este análisis estadístico, donde lo que nos interesa es conocer qué variables pueden suponer un riesgo de cara al desarrollo de cáncer de mama. Además, debido al carácter retrospectivo de la investigación, la OR es la medida idónea para establecer la relación que existe entre el cáncer de mama y los factores o variables de exposición sometidas a estudio.

1.3. El modelo logístico

En esta sección profundizamos un poco más en la odds y la odds ratio dentro del contexto de la regresión logística. Además, trataremos de forma breve los métodos estadísticos utilizados para la estimación de parámetros y los intervalos de confianza para la odds y la odds ratio, el test de Wald para comprobar la significación estadística de los coeficientes, el procedimiento de selección de variables, diferentes test de bondad de ajuste de los modelos, y cómo se puede evaluar su capacidad diagnóstica o predictiva.

1.3.1. La función logit

De acuerdo con lo comentado previamente, el modelo de regresión logística se utiliza en el estudio de la probabilidad de éxito de un suceso de interés, así como su dependencia a las variables explicativas tenidas en cuenta. En el caso que nos ocupa, la probabilidad de éxito, es decir, $P(Y = 1|X = x)$ se corresponde con las probabilidades de sufrir cáncer de mama cuando $X = x$, y la probabilidad de fracaso, $P(Y = 0|X = x)$, a la probabilidad de no sufrir cáncer de mama cuando $X = x$. Para medir la probabilidad de éxito, la regresión logística utiliza medidas de efecto como la odds y la odds ratio. Como hemos visto en apartados anteriores, la odds se define como un cociente entre el número de individuos que presentan una determinada condición, en este caso, una enfermedad (éxito), y el número de individuos que no la padecen (fracaso).

La OR, por otro lado, es una razón o cociente de odds de un grupo de sujetos (enfermos y no enfermos) expuestos, frente a la odds de sujetos (enfermos y no enfermos) no expuestos. Es una medida de asociación entre dos variables que, tal y como hemos comentado previamente, indica la fortaleza y la dirección de la relación que existe entre ambas variables (Cárdenas, 2015), y que se define bajo el modelo logístico como

$$OR = \frac{P(Y{=}1|X{=}x_i)/P(Y{=}0|X{=}x_i)}{P(Y{=}1|X{=}x_j)/P(Y{=}0|X{=}x_j)},$$

donde el numerador se corresponde con la odds cuando $X = x_i$ y el denominador, con la odds cuando $X = x_j$. En epidemiología, es habitual que $X = x_i$ se refiera al grupo de individuos expuestos a un factor de riesgo, y $X = x_j$, al grupo de no expuestos. Dada esta definición, la OR parece idónea para comparar la influencia de uno o varios factores de exposición sobre la aparición de una enfermedad bajo un modelo de regresión logístico. Tal y como veremos a continuación, el modelo logístico puede ser definido como un modelo lineal para el logaritmo de la odds. Sea $\pi(x, \beta) = P(Y = 1|X = x)$ la probabilidad de éxito condicionada para $X = x$, y $\beta = (\beta_0, \beta_1, \cdots, \beta_k)$ el vector de parámetros del modelo logístico que debemos estimar. Bajo estas condiciones, la odds podría definirse como

$$\frac{\pi(x,\beta)}{1-\pi(x,\beta)}.$$

La probabilidad de éxito y la odds están, por lo tanto, relacionadas, pudiendo conocer el valor de una de ellos a partir de la otra. Sin embargo, debemos tener en cuenta que estos estimadores adoptan valores ajustados a escalas diferentes. Mientras la probabilidad de éxito solo puede tomar valores en el intervalo [0,1], la odds lo hace en el intervalo $[0, +\infty]$. Si queremos establecer un modelo de regresión capaz de explicar la relación existente entre la probabilidad de éxito y de fracaso de forma lineal, debemos modificar el soporte de la odds, esto es, permitir que sus valores se muevan en el intervalo $[-\infty, +\infty]$. Para ello, es necesario considerar una función *logit* dada por

$$g(p) = \log \tfrac{p}{1-p} \quad \forall p \in [0,1],$$

mediante la cual considerar un modelo que exprese el logaritmo de la odds como función lineal de las variables explicativas de la forma

$$\log \tfrac{\pi(x,\beta)}{1-\pi(x,\beta)} = x'\beta.$$

De este modo, podemos escribir $g(\pi(x,\beta)) = x^{'}\beta$ (Hosmer et al., 2013), o lo que es lo mismo:

$$g(\pi(x,\beta)) = \log \tfrac{\pi(x,\beta)}{1-\pi(x,\beta)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Además, considerando la inversa de la función *logit*, podemos deducir que

$$\pi(x,\beta) = g^{-1}(x'\beta) = \tfrac{e^{x'\beta}}{1+e^{x'\beta}}.$$

De esta manera, interpretar los resultados resulta más sencillo dado que la odds y la OR son medidas estandarizadas, libres ante cambios de escala de las variables incluidas en el modelo (Cárdenas, 2015). En la Figura 1.3, hemos ajustado un modelo logístico para la base de datos de cáncer de mama, considerando como variable explicativa el NLR. El ajuste se adapta mejor al comportamiento de los datos que un ajuste lineal, presentado en la Figura 1.2. Vemos que bajo un ajuste logístico, la cantidad de casos aumenta a medida que lo hace el NLR, por lo que podríamos estudiar si esta variable supone un factor de riesgo real del cáncer de mama.

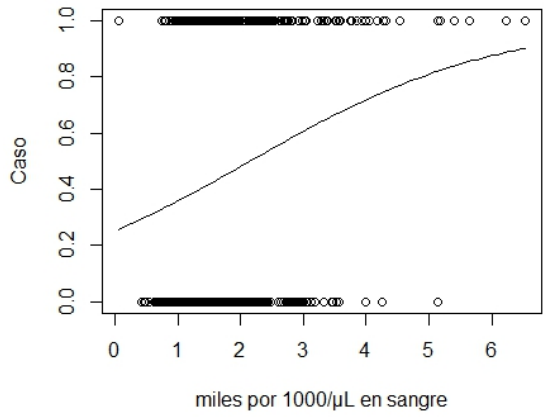


Figura 1.3: Ajuste del modelo de regresión logístico con variable respuesta igual a la variable indicadora de cáncer de mama y la ratio neutrófilos/linfocitos como variable explicativa .

En particular, si X es también una variable dicotómica y solo toma los valores 0 y 1, indicando por ejemplo, la exposición (1) o no (0) a un factor de riesgo, la OR vendría dada por

$$OR = \frac{\frac{\pi(1,(\beta_0,\beta_1))}{[1-\pi(1,(\beta_0,\beta_1))]}}{\frac{\pi(0,(\beta_0,\beta_1))}{[1-\pi(0,(\beta_0,\beta_1))]}},$$

Por tanto, para este modelo de regresión logística en particular, obtendríamos

$$OR = \frac{\frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}}{\frac{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\left(\frac{1}{1+e^{\beta_0}}\right)}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}.$$

Así, si por ejemplo nuestra variable respuesta dicotómica es la presencia o ausencia de cáncer de mama, y nuestra variable explicativa también dicotómica es el consumo o no de anticonceptivos orales, una OR = 2 significa que la odds de padecer cáncer de mama es el doble en las sujetos que han tomado la píldora en comparación a la odds de las que no la han tomado.

Si X es una variable continua, la interpretación de los coeficientes depende de las unidades particulares de la variable (Hosmer et al., 2013). En este caso,

$$Odds(Y|X = x_0) = \frac{e^{\beta_0+\beta_1 x_0}/(1+e^{\beta_0+\beta_1 x_0})}{1-e^{\beta_0+\beta_1 x_0}/(1+e^{\beta_0+\beta_1 x_0})} = e^{\beta_0+\beta_1 x_0}$$

y

$$OR=\frac{e^{\beta_0+\beta_1(x_0+1)}}{e^{\beta_0+\beta_1 x_0}} = e^{\beta_1}$$

Así, $OR = e^{\beta_1}$, representa la odds ratio del incremento en la variable explicativa una unidad (Szumilas, 2010). Por lo tanto, si por ejemplo nuestra variable respuesta dicotómica es la presencia o ausencia de cáncer de mama, y nuestra variable explicativa continua es la edad, si la $OR = 3$ esto significa que por cada año cumplido, la odds de desarrollar cáncer de mama se triplica.

1.3.2. Estimación de parámetros

Para estimar el vector de parámetros β del modelo logístico a partir de una muestra aleatoria simple $(X_1, Y_1), \ldots, (X_n, Y_n)$ de tamaño n del par (X, Y), suele utilizarse la estimación por máxima verosimilitud, maximizando la conocida como **función de verosimilitud**:

$$L(\beta) = \prod_{i=1}^{n} [\pi(x_i, \beta)^{y_i} (1 - \pi(x_i, \beta))^{1-y_i}],$$

cuyo logaritmo es:

$$\log L(\beta) = \sum_{i=1}^{n} [y_i \log \pi(x_i, \beta) + (1 - y_i) \log (1 - \pi(x_i, \beta))],$$

donde $\pi(x_i, \beta)$ y $1 - \pi(x_i, \beta)$, son las probabilidades de éxito y fracaso condicionadas a un valor de $X = x_i$, respectivamente (Hosmer et al., 2013). Para hallar una solución a estas ecuaciones, debemos recurrir a la derivada parcial respecto del parámetro β, que bajo un contexto de regresión logística se expresan de la siguiente manera para los dos casos anteriores:

$$\frac{\partial \pi(x, \beta)}{\partial \beta} = x' \pi(x, \beta)(1 - \pi(x, \beta)),$$

y

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i' [y_i - \pi(x_i, \beta)] = 0.$$

Éstas son conocidas como **ecuaciones de verosimilitud**, que serán usadas para obtener el estimador $\hat{\beta}$ de β. Sin embargo, estas ecuaciones no poseen una solución explícita, por lo que es habitual el uso de métodos iterativos como el de Newton-Raphson o el IRLS (*Iteratively Reweighted Least Squares*) para la estimación de los vectores de parámetros de los modelos ajustados. El software RStudio utiliza IRLS, por eso será la aproximación considerada en este trabajo.

1.3.3. Intervalos y contrastes para la odds y la odds ratio

En el contexto de la regresión logística, es común el uso del método **profile likelihood** o **perfil de verosimilitud** para el cálculo de intervalos de confianza (IC) para el vector de parámetros β. El perfil de verosimilitud de un parámetro β_j, es igual a:

$$PL(\beta_j) = \max_{\beta_1,\ldots,\beta_{j-1},\beta_j,\beta_{j+1},\ldots,\beta_p} L(\beta_1,\ldots,\beta_{j-1},\beta_j,\beta_{j+1},\ldots,\beta_p),$$

una expresión que fija un valor para β_j y maximiza la función de verosimilitud con respecto al resto de parámetros del modelo. Sin embargo, fijar de esta manera el valor de un parámetro minimiza la verosimilitud, por lo que se seleccionará el valor de β_j que, a pesar de ello, sea más verosímil. Los valores de este parámetro que mayor verosimilitud demuestren, formarán parte del intervalo de confianza para β_j. De este modo, se establece el IC en base a los valores de β_j más verosímiles a un nivel de confianza $1-\alpha$ como:

$$\left\{\beta_j : 2(PL(\hat{\beta}_j) - PL(\beta_j)) < \chi^2_{1,\alpha}\right\}$$

siendo $\chi^2_{1,\alpha}$ el cuantil $(1-\alpha)$ de la distribución ji-cuadrado con un grado de libertad. En este trabajo, utilizaremos este método para el cálculo de los IC de los parámetros de los modelos de regresión logística ajustados, así como para sus odds y odds ratio, pues como hemos visto, las estimaciones del vector $\hat{\beta}_j$ pueden ser interpretados como odds y odds ratio aplicando la función exponencial a sus valores. Por lo tanto, una vez calculados los IC para los parámetros β, podremos conocer los IC de la odds y la odds ratio aplicando la función exponencial a los extremos de los intervalos.

Además de este método, pueden estimarse los intervalos de confianza para β_j de manera asintótica, utilizando la aproximación de la distribución del estimador a través de una distribución normal. De esta forma, los IC podrían definirse de la siguiente manera:

$$(\hat{\beta}_j - z_{\alpha/2}\hat{\sigma}(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\hat{\sigma}(\hat{\beta}_j)),$$

donde $z_{\alpha/2}$ se corresponde con el cuantil $(1-\alpha/2)$ de la distribución normal estándar, y $\hat{\sigma}$ es la desviación típica estimada de $\hat{\beta}_j$. Con muestras de tamaño pequeño o moderado, no es recomendable utilizar este método, pues se han visto comportamientos asimétricos de los estimadores $\hat{\beta}_j$ y por lo tanto, de sus respectivos intervalos de confianza.

1.3.4. Contraste sobre los parámetros del modelo

El **test de Wald** permite contrastar la hipótesis nula $\beta_j = 0$, es decir, nos indica si podemos prescindir de un parámetro en el ajuste del modelo de regresión logística. Es necesario considerar como estadístico de contraste

$$\frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)},$$

que sigue una distribución $N(0,1)$ bajo la hipótesis nula cuando el tamaño de muestra n es grande. Como vemos, no es más que una ratio entre el estimador de máxima verosimilitud $\hat{\beta}_j$, y la estimación de su desviación típica, $\hat{\sigma}(\hat{\beta}_j)$ (Hosmer et al., 2013).

1.3.5. Contraste de modelos mediante la *deviance*

La **deviance** de un modelo es un estadístico de razón de verosimilitudes del modelo de interés frente al modelo saturado, un modelo hipotético cuyas predicciones de Y coinciden con los valores observados en la muestra de estudio. En el caso de la regresión logística, la *deviance* se puede escribir como:

$$D_{Modelo} = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\pi(x_i,\hat{\beta}_{Modelo})}{y_i}\right) + (1-y_i)\log\left(\frac{1-\pi(x_i,\hat{\beta}_{Modelo})}{1-y_i}\right)\right].$$

Es común utilizar la deviance como una medida de ajuste similar al coeficiente de determinación de los modelos de regresión lineal en el contexto logístico (Sheater, 2009). Para contrastar dos modelos, donde uno de ellos (hipótesis nula, H_0) es una simplificación del otro (hipótesis alternativa, H_a), se usa el test de razón de verosimilitudes basado en la *deviance*, cuyo estadístico de contraste coincide con la diferencia de *deviance* de ambos modelos ($DH_0 - DH_a$), que sigue una distribución ji-cuadrado.

1.3.6. Bondad de ajuste del modelo

En regresión lineal, el coeficiente de variación o R^2, resume la proporción de varianza en la variable respuesta asociada con las variables explicativas, midiendo así la capacidad predictiva del modelo de regresión y la fuerza de la relación entre las variables que lo conforman. Sin embargo, necesita ser adaptado al contexto logístico, pues este tipo de modelos de regresión no son homocedásticos y por lo tanto, su varianza no es constante. En J. Smith y M. McKenna (2013), se ponen a prueba diferentes adaptaciones del coeficiente de determinación, que al no cumplir las mismas especificaciones que el R^2 clásico, son llamados pseudo R^2.

Uno de los más conocidos es el pseudo R^2 de **MacFadden**, que compara un modelo sin ninguna variable predictora, esto es, considerando únicamente la variable respuesta, frente a un modelo con todas las covariables (Sheater, 2009). Sin embargo, para Smith y McKenna (2013) la bondad de ajuste se mide mejor con el pseudo R^2 de **Aldrich-Nelson**, utilizando la corrección de **Veall-Zimmermann**, y el de **Nagelkerke**, que funcionan de forma similar a McFadden pero cuentan con reescalados o correcciones.

A pesar de la utilidad de estas pruebas, Smith y McKenna (2013) apuntan que los resultados que arrojan no suelen ser tan exactos como el R^2 de la regresión lineal, y que pueden ser susceptibles a cambios en el tamaño muestral o en la naturaleza de las variables (continuas o discretas). Por esto, es recomendable acompañarlos con pruebas de bondad de ajuste como el **test de Hosmer-Lemeshow**, fundamentado en la prueba χ^2 de Pearson, y que consiste en comparar los valores predichos por el modelo con los observados. Bajo la hipótesis nula, no existen diferencias entre los valores observados y los predichos, por lo que si un modelo no supera el test, esto significaría que no está bien ajustado y que las variables explicativas no explican el comportamiento de Y. Haremos uso de todas estas pruebas para estudiar la bondad de ajuste de los modelos logísticos ajustados en este trabajo.

1.3.7. Selección de variables

Cuando el número de variables explicativas disponibles es muy alto, como sucede en nuestro caso, son varios los modelos posibles como resultado de la combinación de estas variables. Por ello, podemos hacer uso de criterios como el AIC (*Akaike Information Criterion*), una medida estadística definida como:

$$AIC = -2 \times L + 2 \times (k + 1),$$

siendo L el logaritmo de verosimilitud del modelo y $k + 1$ el número de coeficientes de regresión estimados, determinados por el número de variables explicativas (Hosmer et al., 2013). Generalmente, cuando se comparan distintos modelos, aquel con el AIC más pequeño es reconocido como el mejor, pues se supone el modelo más equilibrado en cuando a verosimilitud y número de parámetros.

1.3.8. Diagnosis del modelo logístico

Los modelos de regresión logística deben satisfacer una serie de suposiciones que han sido asumidas para su planteamiento (Lawrence, 2006 y Stoltzfus, 2011). Además de la asunción de que la variable respuesta es binaria, debe existir linealidad entre cada variable explicativa continua y la transformación logarítmica aplicada por la función *logit* sobre la variable respuesta. Para comprobar que se cumple esta condición, podemos representar gráficamente las observaciones de cada variable continua en el eje de ordenadas frente a los valores ajustados en el eje de abscisas, y observar si existe una relación lineal entre ellas.

También es necesario comprobar que no esté presente el fenómeno conocido como *colinealidad*. Este ocurre cuando dos o más variables explicativas presentan mucha correlación entre ellas, lo que significa que sus observaciones no son independientes entre sí. Esto provoca que el efecto que las variables explicativas presentan sobre la variable respuesta sea difuso y pueda dar lugar a estimaciones sesgadas de los coeficientes. Debemos tener en cuenta que si dos o más variables tienen observaciones similares

o fuertemente relacionadas, el efecto que produce cada una de ellas se tendrá en cuenta más de una vez (una por variable) en todas las variables que forman parte del modelo de regresión. Cuando más de dos variables presentan mucha correlación, lo correcto es hablar de un problema de multicolinealidad. Para detectar la presencia de colinealidad, se hace uso de los **factores de inflación** de la varianza (VIF, *Variance Inflation Factor*), una herramienta de medición que utiliza la varianza para determinar en qué medida una variable es influida por otra, analizando su contribución al error estándar de la regresión.

Por último, al igual que en otros modelos de regresión, la ausencia de valores atípicos debe ser constatada utilizando medidas como la **distancia de Cook**.

1.4. Curvas ROC

Los modelos de regresión logística se usan a menudo como herramientas de clasificación dentro del ámbito de la biomedicina. A partir de las probabilidades estimadas con el modelo logístico

$$\hat{\pi}(x,\hat{\beta}) = \frac{e^{x'\hat{\beta}}}{1+e^{x'\hat{\beta}}},$$

pueden establecerse reglas de clasificación. Si $\hat{\pi}(x,\hat{\beta})$ es mayor que un umbral prefijado $c > 0$, el individuo se clasifica como éxito (en nuestro caso, sufrir cáncer de mama); si $\hat{\pi}(x,\hat{\beta})$ es menor o igual a c, el individuo se clasifica como fracaso (en nuestro caso, no sufrir cáncer de mama).

Sin embargo, al igual que otras pruebas o test diagnósticos, es muy frecuente que se cometan errores de clasificación. Por ello, es altamente recomendable el uso de métodos capaces de evaluar la capacidad diagnóstica del modelo logístico. Esto es, su competencia para discriminar entre sujetos que presentan o no la condición de enfermedad. Suelen utilizarse con este fin las curvas ROC, una herramienta estadística gráfica usada para evaluar la capacidad discriminatoria de test diagnósticos binarios (Yang et al., 2017 y Fanjul, 2020).

Por tanto, la regresión logística, como método de clasificación de tipo binario, presenta dos posibles resultados: positivo (P) o negativo (N), dependiendo de las probabilidades obtenidas para un sujeto; y puede cometer errores de dos tipos: falsos positivos (FP) y falsos negativos (FN) (Yang et al., 2017 y Pita et al., 2003). Un FP en el contexto que nos ocupa supondría la clasificación de un sujeto sano como enfermo por el modelo de regresión, esto es, que una mujer sin cáncer de mama sea clasificada como enferma de cáncer de mama. Un FN por lo tanto, ocurre cuando un sujeto enfermo es clasificado como sano por el modelo. En este caso, una mujer con cáncer de mama sería identificada como sana, y por lo tanto, sin cáncer de mama. De forma contraria, los aciertos a la hora de clasificar pueden ser también de dos tipos: verdaderos positivos (VP) y verdaderos negativos (VN). De esta forma, los VP serían todas aquellas mujeres clasificadas como enfermas de cáncer de mama cuando realmente

lo están, y los VN, las mujeres clasificadas como no enfermas de cáncer de mama cuando realmente no lo están. En el Cuadro 1.4, se resumen todos los posibles resultados obtenidos por un método de clasificación binario.

		Condición real	
		Sano (S)	Enfermo (E)
Diagnóstico	Positivo (P)	Falso Positivo (FP)	Verdadero Positivo (VP)
	Negativo (N)	Verdadero Negativo (VN)	Falso Negativo (FN)

Cuadro 1.4: Posibles clasificaciones de una prueba diagnóstica binaria.

Utilizando esta matriz de confusión, podemos obtener la **precisión** del modelo al clasificar como

$$Precisión = \frac{\#VP+\#VN}{\#FP+\#VP+\#VN+\#FN}.$$

Además, de cara a la clasificación de los sujetos, es necesario seleccionar el punto de corte c que nos permita realizar la clasificación, del cual hablaremos más adelante. Una vez establecido este umbral, podemos medir la exactitud diagnóstica de una prueba mediante los conceptos de sensibilidad y especificidad. La **sensibilidad** se define como la probabilidad de que la prueba haya dado positivo cuando el sujeto está realmente enfermo, o lo que es lo mismo, la probabilidad de verdadero positivo. En regresión logística, se calcula como

$$Sensibilidad(c) = P(\hat{\pi}(x, \hat{\beta}) > c \mid Y = 1),$$

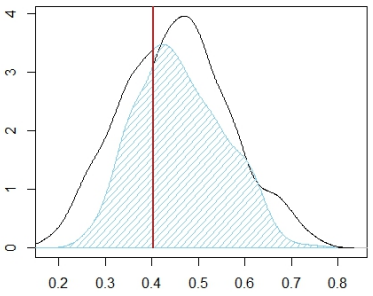
donde c denota un umbral de probabilidad.

La **especificidad** por otro lado, mide la probabilidad de que la prueba haya dado negativo cuando el sujeto está realmente sano, es decir, la probabilidad de verdadero negativo. Bajo los supuestos del modelo de regresión logística, se calcula como

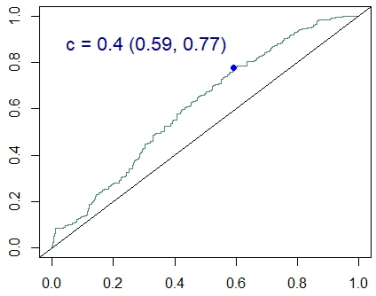
$$Especificidad(c) = P(\hat{\pi}(x, \hat{\beta}) \leq c \mid Y = 0).$$

Nótese que, tanto especificidad como sensibilidad varían dependiendo del punto de corte c escogido. Su papel es clave en la construcción de curvas ROC, obtenidas enfrentado de forma gráfica la sensibilidad (eje Y) con los falsos positivos (eje X), para todos los posibles valores de c. En la curva ROC podremos observar la proporción de falsos positivos frente a la proporción de verdaderos positivos para cada uno de los umbrales posibles, que en nuestro caso serían todo el rango de probabilidades desde 0 hasta 1.

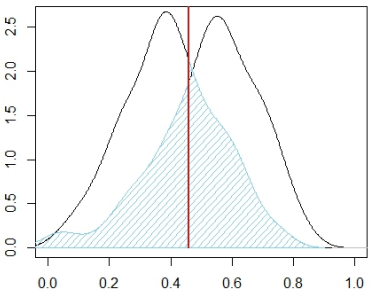
En el caso del modelo logístico, la curva ROC tendrá un comportamiento diferente dependiendo de las funciones de densidad asociadas a las probabilidades del grupo de los enfermos y de los sanos. Si observamos la Figura 1.4, podemos estudiar los cambios que sufre esta curva dependiendo del solapamiento que presenten las funciones de densidad de $\hat{\pi}(x, \hat{\beta})$ para ambos grupos. Las densidades presentes en (a) responden a las probabilidades de un modelo ajustado con las variables independientes NLR, neutrófilos en sangre bajo valores absolutos, y neutrófilos en sangre bajo valores porcentuales. Las densidades en (c) se corresponde con el ajuste de las variables H_2O_2 y edad. Por último, las densidades que podemos ver en (e), son fruto del ajuste de las variables edad, meses dando el pecho, uso de anticonceptivos orales, H_2O_2, antecedentes familiares de cáncer de mama y/o ovárico, NLR y número de hijos.



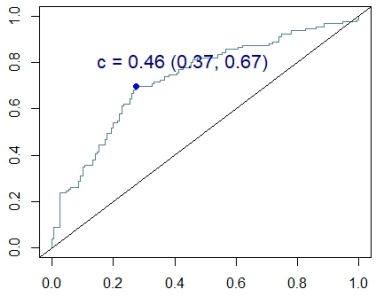
(a) Solapamiento alto



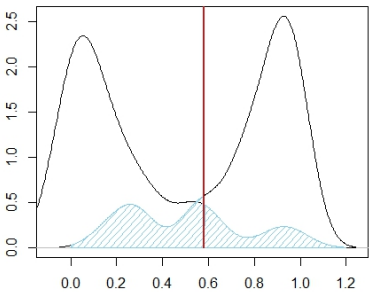
(b) Curva ROC asociada a (a)



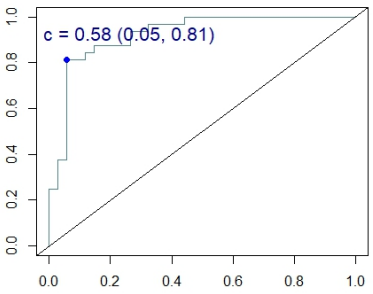
(c) Solapamiento medio



(d) Curva ROC asociada a (c)



(e) Solapamiento bajo



(f) Curva ROC asociada a (e)

Figura 1.4: Relación entre la curva ROC y la función de densidad de $\hat{\pi}(x, \hat{\beta})$ en el grupo de sanos y enfermos.

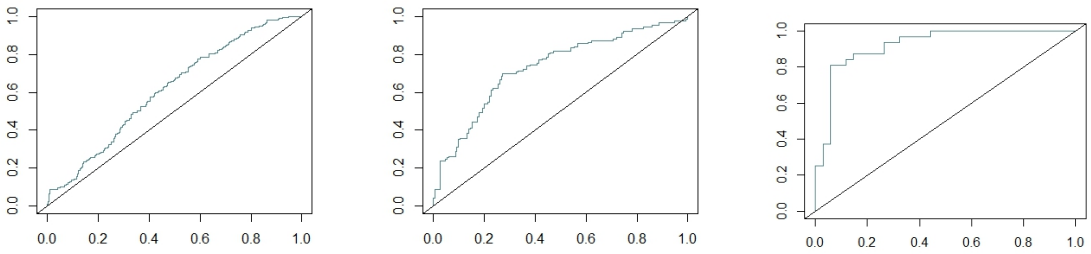
Se aprecia claramente que cuando ambas densidades son muy parecidas, como en el caso de (a), la prueba diagnóstica tiene peor comportamiento, pues vemos en (b) que la curva ROC asociada a este escenario se aproxima mucho a la diagonal, lo que significa que la prueba comete un porcentaje de error muy elevado, siendo muy alta la proporción de falsos positivos (0.59) por cada verdadero positivo (0.77). Cuanto mayor sea la diferencia de probabilidades predichas para el grupo de sanos y de enfermos, más aciertos se darán a la hora de clasificar. La curva ROC (d) asociada a (c) presenta una proporción de verdaderos positivos mayor a la curva anterior, siendo este valor de 0.67, mientras la cantidad de falsos positivos disminuye a un 0.37. La curva (f) asociada a (e) muestra una mayor exactitud al clasificar, con una proporción de verdaderos positivos del 0.81 y de falsos positivos del 0.05. En este caso, las densidades de probabilidad de sanos y enfermos presentan un solapamiento mucho menor, por lo que es más fácil distinguir entre ambos grupos.

Asociadas a la curva ROC, existen medidas resumen de la capacidad discriminatoria de una variable diagnóstico, siendo en regresión logística las predicciones de probabilidad $\hat{\pi}(x, \hat{\beta})$. En nuestro caso, utilizaremos el **AUC** (*Area Under the Curve*), un índice muy útil para comparar el comportamiento de curvas ROC, y es definida habitualmente como:

$$\text{AUC} = \int_0^1 ROC(t)dt$$

Su valor oscila entre 0.5 y 1, informando de la habilidad de la variable o variables diagnósticas escogidas para discriminar entre sujetos sanos y enfermos, en todos los posibles puntos de corte. Un AUC = 0.5 significa que el modelo no es capaz de clasificar de la forma correcta, identificando a un sujeto como sano o enfermo con un 0.5 de probabilidades. Cuanto más se acerque su valor a 1, mejor será el modelo a la hora de clasificar observaciones (Fanjul Hevia, 2020). En la Figura 1.5 podemos

ver los valores del área bajo la curva correspondientes a cada una de las curvas ROC de la Figura 1.4. Cuanto menor solapamiento presenten las densidades de probabilidad $\hat{\pi}(x, \hat{\beta})$, mayor será el valor del AUC asociado a la curva ROC, siendo su máximo el 1. En (a), donde la curva ROC representada se acerca mucho a la diagonal, el valor del área bajo la curva es 0.61, un valor próximo a 0.5. A medida que la curva se aproxima a la esquina superior izquierda, el valor del AUC se aproxima a su vez a 1. Así, el área bajo la curva de (c) es muy superior a las demás, con un valor de 0.9219.



(a) AUC = 0.61 (b) AUC = 0.73 (c) AUC = 0.92

Figura 1.5: Relación entre la curva ROC y el valor del AUC.

Para el caso de la regresión logística, suele utilizarse el estadístico U de Mann-Whitney para la estimación del AUC, dando como resultado la siguiente expresión:

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{j \in D_0} \sum_{k \in D_1} [I(\hat{\pi}(x_j, \hat{\beta}) < \hat{\pi}(x_k, \hat{\beta})) + 0.5 I(\hat{\pi}(x_j, \hat{\beta}) = \hat{\pi}(x_k, \hat{\beta}))],$$

donde D_0 y D_1 son los conjuntos de índices para la ausencia o presencia de la enfermedad respectivamente, y n_0 y n_1, los respectivos tamaños muestrales del grupo de sanos y enfermos (Ipaguirre et al., 2019). Bajo esta expresión, la estimación del AUC utiliza todas las observaciones de la muestra, las cuales han sido previamente usadas para el ajuste del modelo de regresión, dando como resultado lo que se conoce como *AUC aparente*. Este suele adquirir valores demasiado optimistas, por lo que es recomendable dividir la muestra en dos partes para así obtener datos de entrenamiento con los que desarrollar el modelo de regresión, y datos de test con los que probar su validez. Dado que nuestra muestra no es muy grande y contiene muchos valores perdidos (tal y como veremos en el Capítulo 2), no es posible separarla en dos, por lo que hemos recurrido al algoritmo de la validación cruzada, que excluye una observación cada vez que se ejecuta, para corregir la estimación del AUC. Bajo este método, se omite una observación de la muestra original y se ajusta el modelo de regresión logístico a partir del resto de observaciones. El modelo resultante se utiliza entonces sobre la observación omitida para el cálculo de sus probabilidades. Este procedimiento se repite n veces, tantas como observaciones

tenga la muestra, excluyendo una observación diferente cada vez (Ipaguirre et al., 2019). En definitiva, el cálculo del AUC corregido mediante el método de validación cruzada, se fundamenta en las estimaciones de probabilidad de todos los individuos de la muestra.

Como hemos dicho, la sensibilidad y la especificidad dependen del umbral de probabilidad c seleccionado. La selección del valor óptimo de c en la práctica es un problema ampliamente estudiado en la literatura. Para obtener el c óptimo en todos estos casos, hemos utilizado el **índice de Youden** (IY), una medida usada para determinar la capacidad discriminatoria de un test, y que se define como

$$IY = sensibilidad + especificidad - 1.$$

Este índice oscila entre [0,1], puesto que las proporciones de VP, VN, FP y FN obtienen valores en este intervalo. De esta forma, cuanto más se aproxime a 1 el índice de Youden, menor será la cantidad de falsos positivos y falsos negativos, sinónimo de una buena capacidad discriminatoria por parte del test. Para la elección del punto de corte óptimo, utilizamos el c que maximiza el valor del índice de Youden de la siguiente manera:

$$IY = \max_{c} |sensibilidad(c) + especificidad(c) - 1|.$$

Así, el umbral escogido será aquel que presente una sensibilidad y especificidad más altas de manera conjunta, aunque no de forma individual. Dependiendo de la enfermedad a diagnosticar, puede que este método no sea el mejor cuando una de ambas medidas de precisión tiene más peso que la otra al clasificar. Para este trabajo utilizaremos este método para la obtención del umbral. En la Figura 1.4, los gráficos de las curvas ROC contienen el c óptimo para cada caso y la proporción de falsos positivos y verdaderos positivos entre paréntesis. Si nos fijamos, en el caso (b), un c óptimo igual a 0.4 nos deja con una proporción de falsos positivos de un 0.59 frente a un 0.77 de verdaderos positivos. El caso (d) cuenta con un c igual a 0.46, con una proporción de falsos positivos del 0.37 y de verdaderos positivos del 0.67. Por último, en el caso (f), debido a las diferencias en las funciones de densidad del grupo de sanos y de enfermos, el c óptimo tiene un valor de 0.58, con una proporción de falsos positivos del 0.05 y de verdaderos positivos del 0.81.

Capítulo 2

Resultados

En este capítulo presentamos los análisis realizados a la muestra facilitada por la Fundación Pública Gallega de Medicina Genómica. El análisis exploratorio de las variables disponibles en la base de datos se presenta en la Sección 2.1. Éstas varialbes responden esencialmente a características antropométricas, sobre el estilo de vida, sobre la concentración de determinadas células y hormonas en sangre, y sobre factores genéticos. La muestra se compone de un total de 672 mujeres, de las cuales 300 son casos (tienen cáncer de mama) y 372 son controles (no tienen cáncer de mama).

La Sección 2.2. contiene el ajuste de modelos de regresión logística múltiple considerados, para determinar qué variables podrían ser de riesgo o de protección, y para predecir las probabilidades individuales de padecer esta enfermedad. También se muestran las curvas ROC obtenidas de cada modelo ajustado para analizar su capacidad discriminatoria.

2.1. Análisis descriptivo

Para empezar la sección, analizaremos primero las variables asociadas a condiciones genéticas de las sujetos de la muestra o a aquellas características sobre las cuales los individuos no tienen poder de decisión. Éstas son la edad de las sujetos, la edad a la menarquia, la edad a la menopausia, el estado menopáusico y los antecedentes familiares de cáncer de mama y/o ovárico. A continuación se presentarán variables referidas al estilo de vida, como el número de hijos, si han dado o no el pecho, el consumo de alcohol por semana, el consumo de tabaco a lo largo de su vida, el uso de terapias hormonales sustitutivas en la menopausia y el uso de anticonceptivos orales. Las variables antropométricas altura (medida en cm), peso (medida en Kg) e IMC (Kg/m^2) también serán tenidas en cuenta, al igual que otras relacionadas con la concentración de determinadas células en sangre como los neutrófilos, los linfocitos, las plaquetas o los monocitos, cuyo recuento se realiza en $10^3/\mu_L$ de sangre.

Se incluyen también análisis de las ratios NLR y PLR (neutrófilos/linfocitos y plaquetas/linfocitos, respectivamente), calculadas a partir del recuento de estas células en $10^3/\mu_L$ de sangre. Variables como los niveles de H_2O_2 y colesterol en sangre, expresadas en n_g/m_L y m_g/d_L respectivamente, también se han analizado. Por último, se incluyen variables asociadas únicamente a los casos, como los subtipos de cáncer presentes en la muestra, su morfología, tamaño, grado y estadio.

2.1.1. Edad y factores genéticos

La Figura 2.1 contiene un histograma (izquierda) y un diagrama de cajas (derecha), para la variable edad. Debemos destacar que esta variable está medida al diagnóstico para los casos, y a la entrevista para los controles. La totalidad de la muestra oscila entre los 24 y los 94 años, con una media de 58 años y una desviación típica de 13.5. Presenta una moda bastante pronunciada alrededor de los 60-65 años, y otra más pequeña entre los 40-50 años. El diagrama de cajas muestra que la media de edad de los controles es mayor en comparación a los casos, siendo de casi 61 para los primeros y de 55.5 para los segundos.

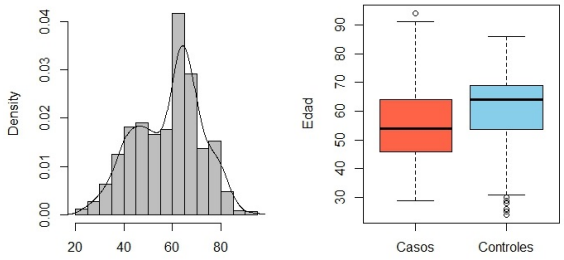


Figura 2.1: Histograma de la edad en la totalidad de la muestra (izquierda) y gráfico de cajas de la edad en base a la variable indicadora de cáncer de mama (derecha).

Si observamos el histograma de la edad separado por casos y controles (ver Figura 2.2), vemos que la edad en los casos presenta una mayor homogeneidad y similitud a la distribución normal, acumulando la mayor parte de las observaciones entre los 45 y los 65 años, edades en las que la incidencia de esta enfermedad aumenta. La edad de los controles, sin embargo, se acumula alrededor de los 60-70 años. La importancia de esta variable en estudios epidemiológicos destaca sobremanera dado que ciertas edades suponen un riesgo en el desarrollo de determinadas enfermedades. La etapa vital de las personas influye en la condición de enfermedad.

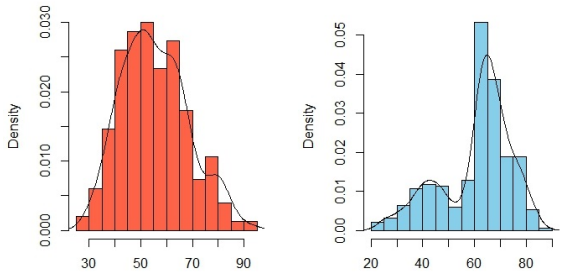


Figura 2.2: Histograma de la edad en casos (izquierda) y en controles (derecha).

Hemos visto que la edad a la menarquia es otra de las variables que pueden ser consideradas de riesgo o de protección (Ritte et al., 2012; Zhang et al., 2015; Phipps et al., 2011; Palmer et al., 2011; Redondo et al., 2012 y Gago-Domínguez et al., 2020). Estudios como Gago-Domínguez et al., (2020), establecen los 13 años como la edad de corte utilizada para clasificar dicha variable como de protección o de riesgo. Una menarquia anterior a los 13 años supondría riesgo, y una posterior, protección. En la Figura 2.3, vemos que la edad a la menarquia de la totalidad de la muestra oscila entre los 9 y los 20 años, con una media de 13.15 y una desviación típica de 1.68. La mayoría de las observaciones se agrupan alrededor de esta media en el histograma, encontrando aquí la moda más pronunciada. Es alrededor de los 12 y los 14 años donde encontramos la mayor densidad de datos. El diagrama de cajas muestra que la media de ambos grupos es muy similar, siendo de 12.97 para los casos y de 13 para los controles. A pesar de estas similitudes, en los histogramas de la densidad de la Figura 2.4 se aprecia un mayor número de casos con menarquias anteriores a los 13 años en comparación a los controles.

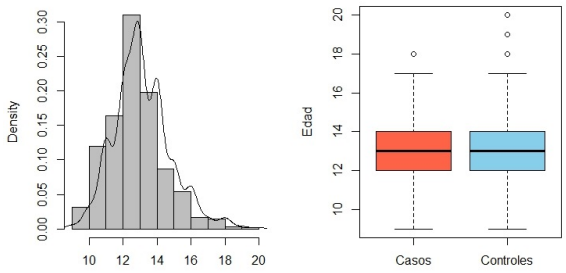


Figura 2.3: Histograma de la edad a la menarquia (izquierda) y gráfico de cajas de la misma variable para casos y controles (derecha).

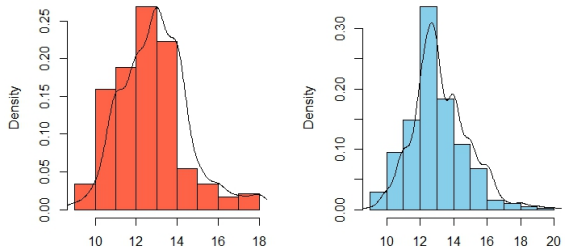


Figura 2.4: Histogramas de la edad a la menarquia para los casos (izquierda) y los controles (derecha).

Al igual que la edad a la menarquia, la edad a la menopausia se supone también una variable de riesgo a medida que esta se pospone en el tiempo (Phipps et al., 2011; Palmer et al., 2011 y Redondo et al., 2012). La muestra cuenta con un mayor número de mujeres diagnosticadas o entrevistadas después de la menopausia, siendo el número de casos con premenopausia ligeramente superior al de controles tal y como podemos observar en el gráfico de mosaicos de la Figura 2.5. Sucede al contrario si observamos los valores postmenopáusicos, donde el número de controles es mucho mayor que el de casos. En el diagrama de cajas de la misma figura, vemos que la media y la desviación típica de la edad a la menopausia es ligeramente superior para los casos, valores que podemos consultar en el Cuadro 2.1.

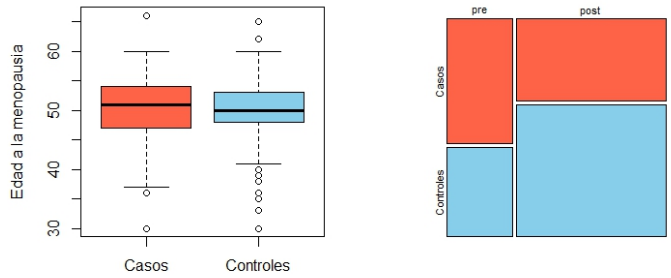


Figura 2.5: Diagrama de cajas de la edad a menopausia en casos y controles (izquierda) y gráfico de mosaico del estado menopáusico (derecha).

Estado menopáusico	pre/post
Muestra al completo	205/467
Casos	120/180
Controles	85/287

Medida de posición/dipersión	M	SD
Muestra al completo	50	5.24
Casos	50.1	5.39
Controles	50	5.15

Cuadro 2.1: Recuento de casos y controles por estado menopáusico (izquierda) y media (M) y desviación típica (SD) muestrales de la edad a la menopasia en la muestra total y por subgrupos según la variable indicadora de cáncer de mama (derecha).

Vemos que la media de edad a la menopausia en casos y controles es muy similar, a pesar de la diferencia en el recuento de casos y controles con menopausia.

En cuanto a los antecedentes familiares, se han tenido en cuenta los casos de cáncer de mama y/o ovárico en uno o más parientes de primer y segundo grado. Observamos tanto en la Figura 2.6 como el Cuadro 2.2, que en el grupo control el número de antecedentes familiares sin cáncer es mayor que en el grupo de casos. Y que, entre los casos, existen más mujeres con antecedentes familiares de cáncer que sin ellos. Esto concuerda con algunos de los estudios que hemos revisado, en los que se hacía una especial mención a esta variable como factor de riesgo del cáncer de mama (ver, por ejemplo, Gago-Domínguez et al., 2016, 2020 y 2021; Jiang et al., 2012; Pharoah et al., 1997; Dolle et al., 2009; Kabat et al., 2012; Ma et al., 2010; Palmer et al., 2011 y Phipps et al., 2011).

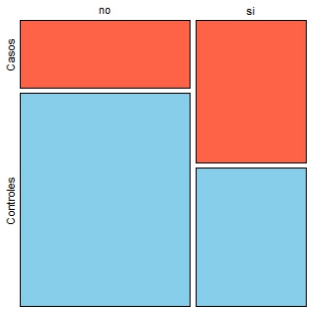


Figura 2.6: Gráfico de mosaicos de los antecedentes familiares para casos y controles.

Antecedentes familiares	si/no/NA
Muestra al completo	158/248/272
Casos	80/58/162
Controles	78/184/110

Cuadro 2.2: Recuento de los antecedentes familiares en la muestra total y por subgrupos de casos y controles. NA es la notación empleada para los datos faltantes.

2.1.2. Variables referidas al estilo de vida

Existen otras variables importantes en el estudio del cáncer de mama relacionadas con decisiones o hábitos de vida. Como vimos en los preliminares, la maternidad, la lactancia o el uso de tratamientos hormonales, pueden alterar o influir en el riesgo de padecer esta enfermedad. Veremos cómo se comporta la muestra con respecto a estas características.

En términos generales, la maternidad es mayor en controles que en casos tal y como aparece recogido en el Cuadro 2.3, y las primeras tienen, de media, un mayor número de hijos (ver Cuadro 2.4). En los gráficos de barras de la Figura 2.7, observamos que los controles con más de 6 hijos son más numerosos que los casos, por lo que podríamos decir que en nuestra muestra esta variable se comporta acorde a la literatura consultada, donde el número de hijos demuestra una relación inversa al riesgo de cáncer de mama y, por lo tanto, se supone como un factor de protección (Phipps et al., 2011 y Palmer et al., 2011).

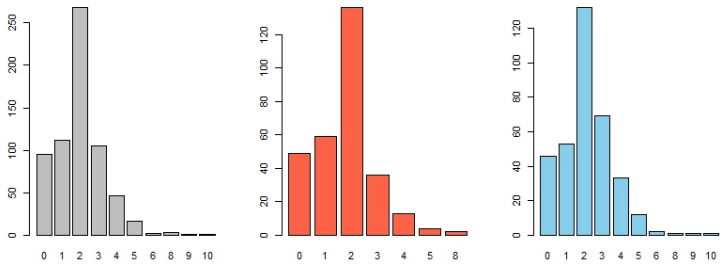


Figura 2.7: Gráfico de barras del número de hijos en la totalidad de la muestra (izquierda), en casos (centro) y en controles (derecha).

En cuanto a la edad al primer hijo nacido, observamos que la media de edad es aproximadamente de 25 años en la totalidad de la muestra, apreciable en el histograma de la Figura 2.8 y en el Cuadro 2.4. Sin embargo, en el gráfico de cajas de esta misma figura vemos que los casos presentan una media de edad ligeramente superior (26.4 frente a y 25, medidas recogidas también en el Cuadro 2.4). Esto significa que los controles tienen, de media, hijos a edades más tempranas que los casos. Sucede a la inversa con el tiempo dedicado a dar el pecho, cuya media es mayor en los controles (6.4), que en los casos (3.72). En el gráfico de cajas de la Figura 2.9 y en el Cuadro 2.4, observamos que efectivamente, los controles sitúan la mayor parte de sus observaciones y la media por encima de las observaciones del grupo de los casos. Además, son más los controles que han hecho uso de esta práctica en comparación a los casos, como podemos ver en el Cuadro 2.3 en relación a la variable Lactancia.

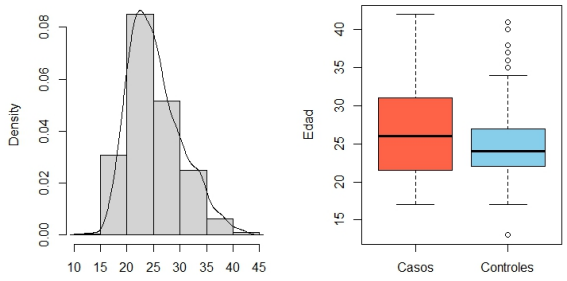


Figura 2.8: Histograma de la edad al primer hijo en la totalidad de la muestra (izquierda) y gráfico de cajas de esta misma variable en casos y controles (derecha).

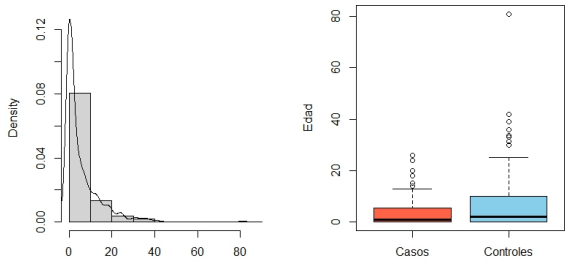


Figura 2.9: Histograma de los meses dando el pecho en la totalidad de la muestra (izquierda), y gráfico de cajas de la misma variable para casos y controles (derecha).

Concluimos así, que el retraso en la edad del primer hijo presenta una relación directa con el riesgo de cáncer de mama, esto es, el retraso en la maternidad aumenta las probabilidades de padecer esta enfermedad. No sucede así con la lactancia, que presenta una relación inversa con el cáncer de mama, por lo tanto, a medida que aumenta el tiempo dando el pecho, disminuyen las probabilidades de desarrollarlo. Como vemos, los resultados de la muestra bajo estudio concuerdan con investigaciones anteriores con respecto a estas variables (Cruz et al., 2013; Redondo et al., 2012; Phipps et al., 2011 y Palmer et al., 2011).

El uso de anticonceptivos orales y terapias hormonales sustitutivas en la menopausia también se ha asociado con un aumento de la incidencia de cáncer de mama (Dolle et al., 2009; Ma et al., 2010; Gago-Domínguez et al., 2020). En la muestra, el uso de anticonceptivos orales es similar en ambos grupos si echamos un vistazo al gráfico de mosaicos de la izquierda de la Figura 2.10 y al Cuadro 2.3. Sin embargo, dentro del grupo control son mucho más numerosas las mujeres que nunca han hecho uso de ellos en comparación al grupo de casos no expuestos. En la misma figura y el mismo cuadro, vemos que el recuento de controles que han utilizado la terapia hormonal sustitutiva es ligeramente superior al de casos, pero también de aquellos que nunca han hecho uso de ella. Cabe destacar que la variable uso de anticonceptivos orales cuenta con una cantidad de valores perdidos (NA) en el grupo de casos muy alta (163 frente a 69 controles). Si añadimos a esto que el número de controles es superior al de casos en la muestra total (300 frente a 372, respectivamente), el resultado es una pérdida de información muy elevada sobre esta variable. Mientras de los controles tenemos un total de 303 observaciones sobre en consumo de estos fármacos, de casos solo 107. Dado el desequilibrio entre ambos grupos, puede que las conclusiones del análisis gráfico sean incorrectas. Bajo la información que tenemos, podríamos suponer que, al igual que en la literatura estudiada, el uso de anticonceptivos orales supone un factor de riesgo porque el número de controles que nunca los han usado es mucho mayor al de casos, y el número de casos es superior en el grupo de expuestos frente al grupo de no

expuestos. Sin embargo, no sabemos qué resultados obtendríamos si tuviésemos a nuestra disposición la misma cantidad de observaciones para casos y controles. La variable Terapia hormonal sustitutiva se encuentra en la misma situación. Bajo el análisis gráfico no parece tener relevancia en el cáncer de mama teniendo en cuenta las características de la muestra, pues no se observan diferencias destacables entre ambos grupos.

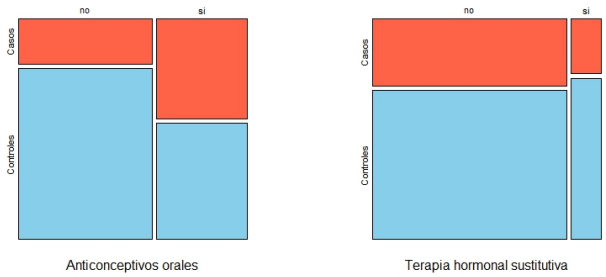


Figura 2.10: Gráficos de mosaico para el uso de anticonceptivos orales (izquierda) y de terapia hormonal sustitutiva (derecha) para casos y controles.

En conductas como el consumo de alcohol, hemos usado las categorías establecidas por el *Centers for Desease Control and Prevention* (CDC, 2022) para lo que consideran como consumo saludable, moderado y de riesgo (0 bebidas alcohólicas a la semana, de 1 a 7, y más de 7, respectivamente). En el gráfico de barras situado a la izquierda en la Figura 2.11, observamos claramente que es mayor el número de controles que hacen un consumo nulo de alcohol, sin embargo, también lo es para aquellas que lo hacen de 1 a 7 veces a la semana, o incluso con una frecuencia mayor a 7. Sucede de forma similar con el consumo de tabaco, siendo mucho más numerosos los controles que nunca han fumado, pero con una cantidad de observaciones de ex y fumadoras prácticamente igual a los casos. Los resultados del Cuadro 2.4 nos indican que la media de consumo de alcohol es superior en controles que en casos, más alta incluso que la media muestral al completo. Numerosos estudios apuntan a que esta práctica aumenta el riesgo de cáncer de mama (Sung et al., 2021; Gago-Domínguez et al., 2016; Ali et al., 2014 y Chen et al., 2011), pero en nuestra muestra los resultados apuntan en otra dirección, pues el consumo moderado y de riesgo en controles es mucho mayor que en casos. Sucede de igual manera para el tabaco. De nuevo, la cantidad de valores perdidos puede entorpecer las conclusiones del análisis.

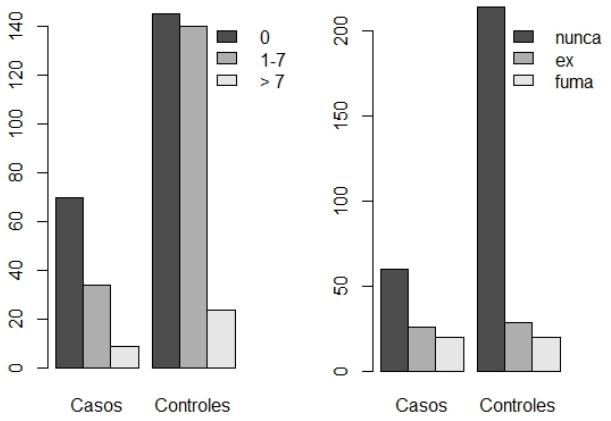


Figura 2.11: Gráfico de barras del consumo de alcohol (izquierda) y tabaco (derecha) para casos y controles.

Variable	**Maternidad**	**Lactancia**	**Anticonceptivos orales**	**Terapia hormonal**	**Tabaco**
Categoría	si/no/NA	si/no/NA	si/no/NA	si/no/NA	nunca/ex/fuma/NA
Muestra total	555/95/22	228/163/281	178/262/232	59/380/233	274/55/40/303
Casos	250/49/1	75/53/172	82/55/163	15/118/167	60/26/20/194
Controles	305/46/21	153/110/109	96/207/69	44/262/66	214/29/20/109

Cuadro 2.3: Recuento de las variables maternidad, lactancia, uso de anticonceptivos orales, terapia hornomal sustitutiva y tabaco en la muestra total y por subgrupos de casos y controles. NA es la notación empleada para los datos faltantes.

Variable	**Nº de hijos**		**Edad primer hijo**		**Meses lactando**		**Consumo alcohol**	
Medida	M	SD	M	SD	M	SD	M	SD
Muestra total	1.98	1.35	25.4	5.02	5.57	8.9	2.89	9.87
Casos	1.76	1.21	26.4	5.91	3.72	5.82	2.34	5.14
Controles	2.16	1.44	25	4.59	6.4	9.87	3.09	6.88

Cuadro 2.4: Media (M) y desviación típica (SD) muestrales de las variables número de hijos, edad al primer hijo, número de meses lacatando y consumo de alcohol a la semana, en la muestra total y por subgrupos de casos y controles.

2.1.3. Variables antropométricas

El peso, la altura y el IMC también parecen tener implicaciones en la aparición y el desarrollo del cáncer de mama, tal y como hemos visto en estudios como Reeves et al. (1996), Cecchini et al.(2012), Kabat et al.(2012) y Zhang et al(2015), muchos de ellos asociados al estado menopáusico (Suzuki et al., 2009; Picon-Ruiz et al., 2017; Trentham-Dietz et al., 1997; y Kawai et al., 2014).

En la muestra de estudio, el peso, medido en Kg, tiene una media superior en casos que en controles, siendo de 69.9 Kg para los primeros y de 68 Kg para los segundos. Pueden consultarse estas medidas en el Cuadro 2.5, junto con la desviación típica y las medias del resto de variables antropométricas analizadas. Vemos en la Figura 2.12, que casos y controles tienen una distribución del peso similar. Sin embago los primeros, presentan observaciones en los valores más altos de la variable si nos fijamos en el gráfico de cajas. Ocurre lo mismo con el IMC (Kg/m^2), aunque la diferencia no es tan grande si observamos la Figura 2.13. En el gráfico de cajas, vemos que los controles se agrupan en un rango de valores ligeramente menor a los casos, pues restando valores atípicos, el resto se concentra en IMC comprendidos entre 20 y 40. Mientras, los casos registran valores comprendidos entre 16 y 42 aproximadamente, lo que sugiere una mayor dispersión de esta variable para este grupo (se puede comprobar en el Cuadro 2.5 que efectivamente la dispersión es menor en el grupo de controles).

En la altura, medida en *cm* tampoco encontramos diferencias reseñables. La media es exactamente la misma para ambos grupos (ver Cuadro 2.5) y su comportamiento en términos de distribución de las observaciones es también muy similar (ver Figura 2.14). En este caso, los resultados de la muestra no

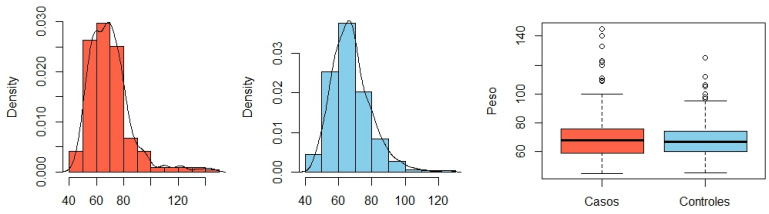


Figura 2.12: Histograma del peso para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

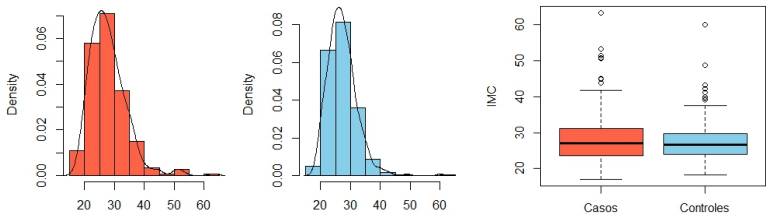


Figura 2.13: Histograma del IMC para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

concuerdan con los estudios revisados, en los que la altura, el peso y el IMC sí suponían un factor de riesgo del cáncer de mama a medida que los valores de estas variables aumentaban.

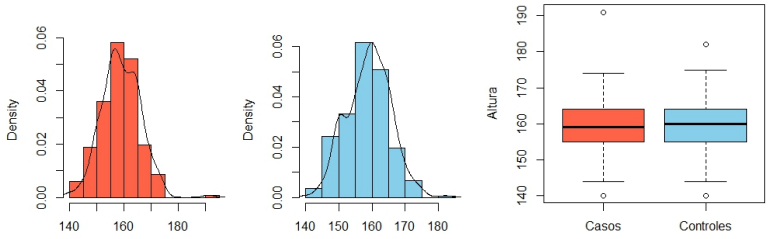


Figura 2.14: Histograma de la altura para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

Variable	**Altura**		**Peso**		**IMC**	
Medida	M	SD	M	SD	M	SD
Muestra total	159	6.76	68.8	13.6	27.6	5.6
Casos	159	7.06	69.9	15.6	28	6.3
Controles	159	6.55	68	11.8	27.3	4.93

Cuadro 2.5: Media (M) y desviación típica (SD) muestrales de las variables altura, peso e Índice de Masa Corporal, en la muestra total y por subgrupos de casos y controles.

2.1.4. Concentración de células sanguíneas

El recuento de glóbulos blancos también ha sido ampliamente estudiado desde la epidemiología del cáncer. En la muestra contamos con el recuento de células como los neutrófilos, los linfocitos, los monocitos y las plaquetas, recogidas en $10^3/\mu_L$ de sangre, y expresadas en variables con valores absolutos y porcentuales (qué proporción ocupa cada una de estas células en la totalidad de glóbulos blancos recogidos en $10^3/\mu_L$).

La distribución de neutrófilos en sangre de la muestra es similar en casos y controles bajo valores absolutos. En el gráfico de cajas de la Figura 2.15 y el Cuadro 2.6, se aprecia que ambos grupos se acercan en media y dispersión. Los controles sitúan un mayor número de observaciones en niveles más bajos que los casos, tal y como podemos comprobar en los histogramas de la misma figura. Las diferencias son más apreciables si medimos el recuento de neutrófilos en proporción. Vemos en el Cuadro 2.7 que las diferencia en media entre casos y controles son mayores, siendo más alta para los primeros. Los casos, por lo tanto, tienen una mayor proporción de neutrófilos en sangre en la totalidad de glóbulos blancos recogidos en $10^3/\mu_L$.

Bajo estos resultados, cabe esperar que una alta concentración de neutrófilos en sangre suponga un factor de riesgo del cáncer de mama. Algunos de los estudios revisados apuntaban hacia estas mismas conclusiones (Uribe-Querol y Rosales, 2015 y Faria et al., 2016), especialmente cuando este recuento es muy superior al nivel de linfocitos en sangre, fenómeno estudiado mediante la ratio neutrófilos/linfocitos o NLR. Si observamos los resultados de la Figura 2.16 y el Cuadro 2.6, vemos que la media de linfocitos bajo valores absolutos es ligeramente superior en controles que en casos, destacando aún

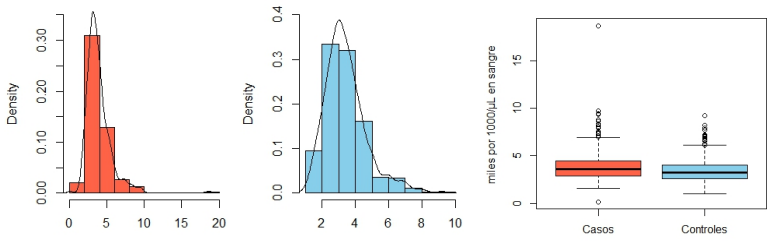


Figura 2.15: Histograma del recuento de neutrófilos en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

más bajo niveles porcentuales (ver Cuadro 2.7). En los histogramas se aprecia una mayor cantidad de observaciones concentradas en niveles altos de linfocitos en el grupo control, observable también en el gráfico de cajas. Al comparar los niveles de NLR en la Figura 2.17, vemos que la media en el gráfico de cajas es ligeramente superior para los casos, como aparece recogido también en el Cuadro 2.8. Además, las observaciones se distribuyen alrededor de valores más altos en el histograma de los casos que en el de los controles. Estos resultados significan que una alta proporción de neutrófilos en comparación a linfocitos en sangre, supone un factor de riesgo en el cáncer de mama. Sin embargo, cuando la cantidad de linfocitos aumenta, las probabilidades disminuyen.

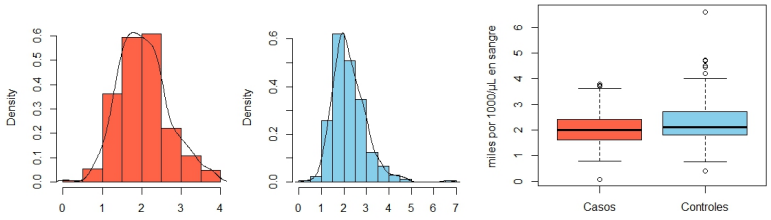


Figura 2.16: Histograma del recuento de linfocitos en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

En la teoría epidemiológica, hemos visto varios estudios que concuerdan en el uso del NLR como un biomarcador de la gravedad y pronóstico del cáncer de mama (Iwase et al., 2016; Faria et al., 2016; Ethier et al., 2017 y Gago-Domínguez et al., 2020), pues niveles altos de neutrófilos con respecto a los de linfocitos parecen indicar un peor pronóstico de la enfermedad. Sin embargo, esto no significa que su presencia sea la causa del cáncer de mama, pues esta ratio sólo informa del futuro de la enfermedad una vez ésta ya ha aparecido. Por ello, los niveles de neutrófilos suelen ser superiores en los casos que

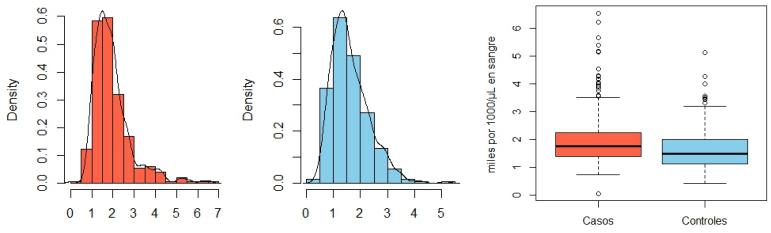


Figura 2.17: Histograma de la ratio NLR en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

en los controles, pues los segundos *a priori*, son sujetos sanos.

Además del recuento de neutrófilos y linfocitos, contamos también con variables como el número de monocitos y plaquetas en sangre, que como vemos en el Cuadro 2.6 tienen una media más alta en casos que en controles, aunque estas diferencias no son muy grandes. En los gráficos de las Figuras 2.18 y 2.19, vemos que la distribución y la densidad de las observaciones es muy parecida en ambos grupos, aunque los monocitos presentan una mayor concentración de casos en los cuantiles más altos de la distribución.

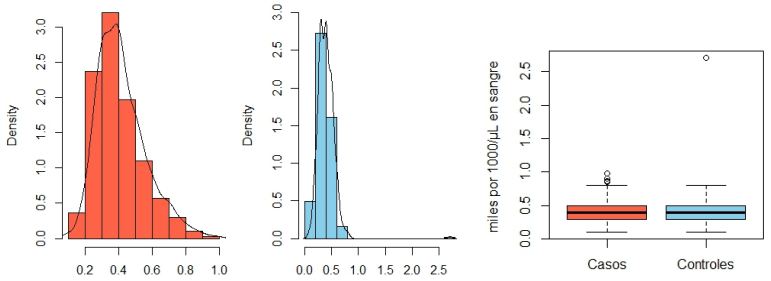


Figura 2.18: Histograma del recuento de monocitos en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

En el Apéndice A podemos consultar los histogramas de densidad y los gráficos de cajas de las variables neutrófilos, linfocitos y monocitos en sangre medidos bajo niveles porcentuales, observando los resultados de las Figuras A.1, A.2 y A.3.

En cuanto a la ratio plaquetas/linfocitos o PLR (*Platelet to Lymphocyte Ratio*), sí encontramos una media superior en casos que en controles, observable en el gráfico de cajas de la Figura 2.20 y en el Cuadro 2.8. Además, los primeros muestran una densidad de datos en valores más altos de la

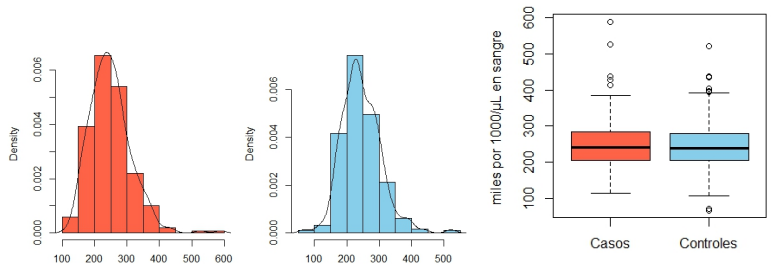


Figura 2.19: Histograma del recuento de plaquetas en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

variable (a partir de los 200), observable en los histogramas de la misma figura, mientras que los controles distribuyen sus observaciones hacia valores más bajos. El PLR ha sido usado también como biomarcador del pronóstico de diferentes cánceres (Zhou et al., 2014) al igual que el NLR, anunciando un peor desarrollo de la enfermedad a las ratios más altas. En la muestra, los resultados coinciden con la literatura consultada, pues tanto el NLR como el PLR tienen concentraciones más altas entre los casos, que aunque no son diferencias excesivamente grandes, podrían ser significativas de cara al futuro de la enfermedad.

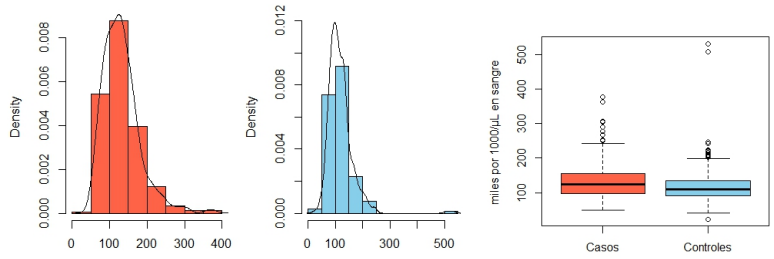


Figura 2.20: Histograma de la ratio PLR en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

Los niveles de H_2O_2 y colesterol en sangre también han sido estudiados. La Figura 2.21 muestra que los casos cuentan con niveles más bajos de H_2O_2 en comparación a los controles. En el gráfico de cajas vemos claramente que la media de los controles es superior, y que las observaciones se distribuyen hacia valores más altos de la variable (ver Cuadro 2.8). En los histogramas de densidad se aprecian estas diferencias entre ambos grupos, donde sin contar con valores atípicos, los controles poseen más observaciones en los cuantiles más altos de la distribución. Contar con una cantidad baja de peróxido

de hidrógeno puede provocar que los efectos del NLR sean más perjudiciales (Gago-Domínguez et al., 2020). Sucede al contrario con el colesterol, que es más alto en media para casos que para controles (ver Cuadro 2.8). No son diferencias muy altas como podemos apreciar en el gráfico de cajas de la Figura 2.22, a pesar presentar una mayor densidad de datos en los valores más altos de la variable en el histograma de los casos. Estudios como Gago-Domínguez et al., (2020), indican que niveles de colesterol altos pueden afectar a la actividad celular en la lucha contra agentes cancerosos, por lo que sus niveles también pueden ser determinantes del pronóstico de la enfermedad. Por los resultados de estos gráficos, concluimos que el H_2O_2 es un factor de protección a medida que sus valores aumentan. Sobre el colesterol sin embargo, no podemos sacar conclusiones concluyentes, pues la distribución de los datos es similar en ambos grupos.

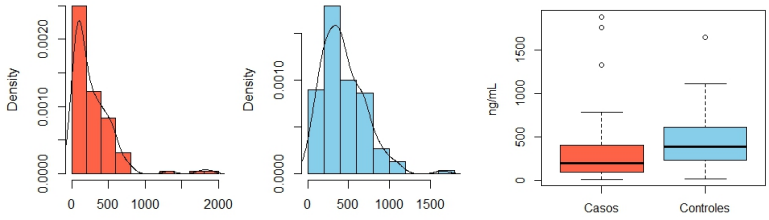


Figura 2.21: Histograma de los niveles de H_2O_2 medidos en n_g/m_L de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

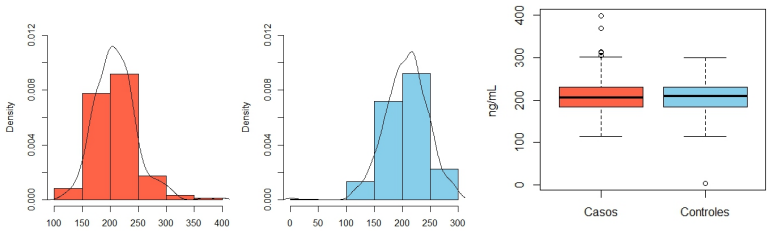


Figura 2.22: Histograma de los niveles de colesterol medidos en m_g/d_L de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

Variable	**Neutrófilos**		**Linfocitos**		**Monocitos**		**Plaquetas**	
Medida	M	SD	M	SD	M	SD	M	SD
Muestra total	3.64	1.51	2.15	0.702	0.41	0.163	247	62.6
Casos	3.88	1.72	2.05	0.636	0.418	0.146	249	64.1
Controles	3.43	1.26	2.25	0.741	0.402	0.176	245	61.3

Cuadro 2.6: Media (M) y desviación típica (SD) muestrales de la concentración de neutrófilos, linfocitos, monocitos y plaquetas en sangre, en la muestra total y por subgrupos de casos y controles.

Variable	**Neutrófilos %**		**Linfocitos %**		**Monocitos %**	
Medida	M	SD	M	SD	M	SD
Muestra total	54.7	9.74	33.6	8.74	6.34	1.73
Casos	57	9.23	31.5	8.2	6.4	1.96
Controles	52.8	9.77	35.4	8.8	6.28	1.51

Cuadro 2.7: Media (M) y desviación típica (SD) muestrales de la concentración de neutrófilos, linfocitos y monocitos bajo niveles porcentuales en $10^3/\mu_L$ de sangre, en la muestra total y por subgrupos de casos y controles.

Variable	**NLR**		**PLR**		**H_2O_2**		**Colesterol**	
Medida	M	SD	M	SD	M	SD	M	SD
Muestra total	1.78	0.828	124	49.2	366	287	208	38.1
Casos	1.96	0.926	132	49.9	282	293	210	38.3
Controles	1.64	0.706	118	47.9	436	264	207	38

Cuadro 2.8: Media (M) y desviación típica (SD) muestrales de la ratio neutrofilos/linfocitos (NLR), plaquetas/linfocitos (PLR) y niveles de H_2O_2 y colesterol en $10^3/\mu_L$ de sangre, en la muestra total y por subgrupos de casos y controles.

2.1.5. Características tumorales de los casos

Ahora, estudiaremos las variables referidas únicamente a los casos: clasificaciones morfológicas, subtipos moleculares, grado y estadio tumoral. Puesto que son variables que sólo atañen a uno de los grupos de la muestra, no serán tenidas en cuenta en el estudio analítico mediante modelos de regresión, pues no aportan información útil con la que explicar el comportamiento de la variable Y.

Los subtipos moleculares de cáncer de mama se construyen en base a los receptores de estrógenos y progesterona en los que se expresan las células cancerosas (DePolo, 2022). El subtipo Luminal A suele ser el más común y el que mejor pronóstico presenta. Es positivo para los receptores de estrógeno y progesterona, y negativo para la proteína HER2, responsable del desarrollo celular. Cuando esta proteína se ve afectada, las células cancerosas se multiplican y diseminan más rápido, que es lo que sucede en el subtipo HER2, resultando en peores pronósticos. El subtipo triple-negativo o TNBC (*Triple Negative Breast Cancer*), presenta receptores negativos de estrógeno, progesterona y HER2. Es el subtipo más agresivo. El Luminal B se diferencia del Luminal A en que presenta receptores de progesterona negativos.

La clasificación morfológica, por otro lado, establece categorías en base al lugar de la mama en el que aparece el tumor. El carcinoma ductal es el más común. Se desarrolla en el revestimiento de los conductos responsables de llevar la leche al pezón. El carcinoma ductal invasivo puede subdividirse en otras categorías dependiendo del modo en que se vean afectadas las células mamarias (medular,

mixto, mucinoso, papilar y tubular) (DePolo, 2023). El carcinoma lobular se origina en las glándulas productoras de leche y es el segundo tipo de cáncer de mama que más se da (DePolo, 2023). El cáncer de mama inflamatorio es poco común, y se manifiesta no como un bulto, sino mediante una sensación de pesadez o espesor en la mama (DePolo, 2023).

Los gráficos de sectores de la Figura 2.23 nos informan de que el subtipo molecular Luminal A es el más frecuente en la muestra, seguido por el TNBC y el Luminal B dentro de la clasificación por receptores afectados. Por morfología, el ductal se manifiesta en el 80 % de los casos, seguido por el lobular (9.34 %).

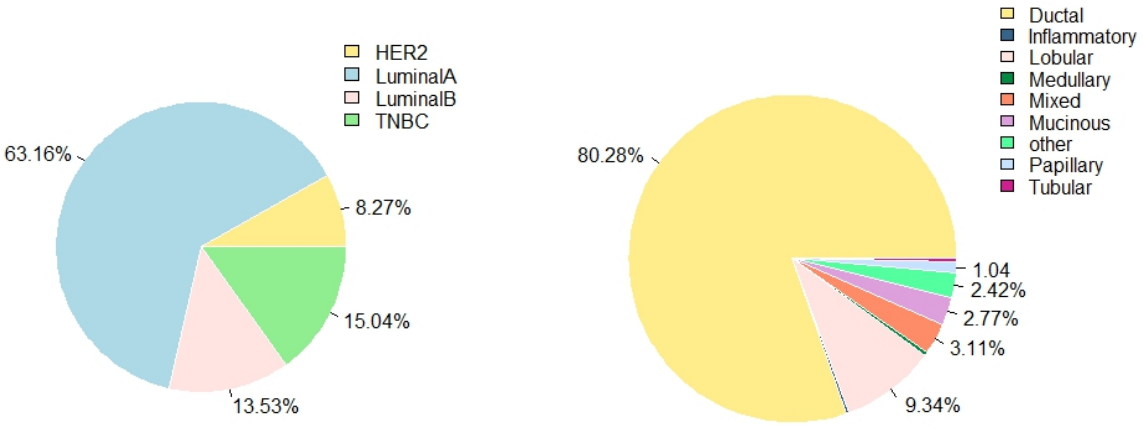


Figura 2.23: Diagrama de sectores por subtipo molecular (izquierda), y por morfología (derecha).

En cuanto al grado tumoral, variable que mide qué tan normales o anormales se ven las células cancerosas al microscopio (Instituto Nacional del Cáncer, 2023), encontramos que la mayoría de la muestra presenta tumores de grado II y III, seguidos por el grado I (130, 73 y 61 casos respectivamente). Cuanto mayor sea el grado tumoral, más agresivo es el cáncer y más rápido crece. El estadio mide la extensión del cáncer teniendo en cuenta el tamaño, el número de nodos o ganglios afectados y la existencia o no de metástasis. En la muestra, el mayor número de casos están en el estadio II, suponiendo el 56.6 % del total de sujetos enfermas. Tanto el estadio como el grado del tumor son variables necesarias para entender la gravedad de la enfermedad, las probabilidades de supervivencia y la planificación del tratamiento. Podemos ver un resumen de todas estas variables en el Cuadro 2.9. En cuanto al tamaño medio de los tumores, los casos muestran una media de 21.2 mm, con una desviación típica de 15.8. Para el número de nodos afectados, obtenemos como valor medio el 1.7, y

una desviación típica de 3.81.

Subtipo morfológico	**Nº de casos(% sobre la totalidad de casos)**
Ductal	232(80.28 %)
Lobular	27(9.34 %)
Inflamatorio	1(0.35 %)
Medular	1(0.35 %)
Mixto	9(3.11 %)
Papilar	3(1.04 %)
Tubular	1(0.35 %)
Otro	7(2.42 %)
Subtipo molecular	
Luminal A	158(63.16 %)
Luminal B	36(13.53 %)
HER2	22(8.27 %)
TNBC	40(15.04 %)
Estadio	
I	85(40.09 %)

II	120(56.6 %)
III	7(3.3 %)
Grado	
I	61(22.51 %)
II	130(47.67 %)
III	73(26.94 %)

Cuadro 2.9: Recuento y proporción de casos por subtipo morfológico, subtipo molecular, estadio y grado, sobre el total de sujetos pertenecientes al grupo de casos.

En resumen, el análisis exploratorio realizado indica que variables como la edad, la edad a la menarquia, y el estado menopáusico, tienen relación con el cáncer de mama. El cumplir años supone un factor de protección contra el cáncer de mama, pues la cantidad de casos con edades avanzadas es menor que la de controles, y estos últimos presentan una media de edad mayor. De igual manera, el retraso en la edad a la menarquia también se considera un factor de protección, esto es, cada año cumplido sin la aparición de la primera regla disminuye el riesgo de cáncer. Sucede al contrario con el retraso de la menopausia, pues esta ha resultado ser un factor de protección. Las sujetos con antecedentes familiares de cáncer de mama y/o ovárico presentan mayores probabilidades de sufrir cáncer de mama en comparación a aquellas que no cuentan con estos antecedentes. Con respecto a la maternidad, el haber tenido hijos no parece tener implicaciones en el cáncer de mama bajo los resultados de la muestra. Sin embargo, el número de hijos, una edad temprana al primer hijo, la lactancia, y una mayor cantidad de meses lactando, sí demuestran relaciones inversas con las probabilidades de desarrollar este tipo de cáncer, pues los controles tienen de media, más hijos, a edades más tempranas, y dan el pecho durante más tiempo que los casos. La relación del uso de anticonceptivos orales con el cáncer de mama no está clara debido a la cantidad de valores perdidos, pero los resultados se decantan por clasificarla como una variable de riesgo. La cantidad de controles que nunca han usado estos fármacos son muchos más que los casos, y entre estos, existen más observaciones en el grupo de expuestas que al de no expuestas. Por otro lado, la terapia hormonal sustitutiva no parece tener impacto en el cáncer de mama, al igual que el consumo de alcohol y tabaco.

En cuanto a las variables antropométricas, los resultados del análisis indican que no existen diferencias entre el grupo de casos y de controles. Las medias son muy similares entre ambos, y aunque la variables peso e IMC presentan más observaciones en los cuantiles altos de la distribución de densidad (ver Figuras 2.12 y 2.13), vemos en el gráfico de cajas que se tratan de muy pocas observaciones.

La cantidad de monocitos, plaquetas y colesterol, tampoco presentan muchas diferencias entre grupos. Las disparidades en el recuento de neutrófilos son más notables bajo valores porcentuales que absolutos (ver Figuras A.1 y 2.15), siendo más alto para los casos que para los controles. Por lo tanto, niveles altos de neutrófilos en sangre en comparación a otros glóbulos blancos, aumentan el riesgo de cáncer de mama. Por ello, la variable NLR también presenta valores más altos en el grupo casos. Junto con el PLR, son ambas variables que indican un mal pronóstico de la enfermedad cuando sus valores son muy altos. Como ambas comparan la cantidad de neutrófilos y plaquetas con respecto a la de linfocitos en sangre, cabe esperar que valores altos de linfocitos tengan una relación inversa al riesgo de cáncer de mama. Y efectivamente así es, pues el recuento de linfocitos presenta una media ligeramente superior para los controles tanto en valores absolutos como porcentuales. Por otro lado, el H_2O_2 también parece ser una variable de protección, pues en la muestra los controles cuentan con una media muy superior a los casos (casi el doble), lo que significa que a medida que aumenta el valor de esta variable, el riesgo de cáncer de mama disminuye.

2.2. Ajuste de modelos de regresión logística

Hemos ajustado distintos modelos de regresión logística múltiple para predecir las probabilidades de sufrir cáncer de mama a partir de las variables contenidas en la base de datos facilitada por la Fundación Pública Gallega de Medicina Genómica. Tal y como hemos visto en la introducción de esta memoria, la regresión logística es una buena alternativa cuando contamos con una variable dependiente binaria (en nuestro caso tener o no cáncer de mama), en base a una serie de variables explicativas.

Nuestra variable respuesta Y recibe el nombre de *Caso*, y solo cuenta con los valores 0 y 1 dependiendo de la ausencia o presencia de cáncer de mama. Las variables explicativas tenidas en cuenta para los ajustes de modelos de regresión son: Edad al diagnóstico o entrevista (X_1), edad a la menarquia (X_2), estado menopáusico (X_3), edad a la menopausia (X_4), maternidad (X_5), número de hijos (X_6), edad al primer hijo (X_7), lactancia (X_8), meses lactando (X_9), IMC (X_{10}), consumo de anticonceptivos orales (X_{11}), uso de terapia hormonal sustitutiva (X_{12}), consumo de alcohol (X_{13}), consumo de tabaco (X_{14}), historial familiar de cáncer de mama y/o ovárico (X_{15}), concentración de neutrófilos en valor absoluto (X_{16}) y en valor porcentual (X_{17}), concentración de monocitos en valor absoluto (X_{18}) y en valor porcentual (X_{19}), concentración de linfocitos en valor absoluto (X_{20}) y en valor porcentual (X_{21}), plaquetas en valor absoluto (X_{22}), NLR (X_{23}), PLR (X_{24}), H_2O_2 (X_{25}), altura (X_{26}), peso (X_{27}) y colesterol (X_{28}). Se han suprimido como posibles variables predictoras aquellas referidas únicamente a características de los casos.

Dado que nos encontramos en un contexto epidemiológico, el objetivo de la regresión logística es arrojar información sobre qué variables pueden ser consideradas de riesgo o de protección, y predecir en base a ellas, las probabilidades de desarrollar cáncer de mama. Nuestra primera opción consistía en ajustar un modelo de regresión logística con las 28 variables explicativas X_i ($i = 1, \ldots, 28$) mencionadas, con el objetivo de obtener un único modelo logístico múltiple a partir del cual realizar una selección de variables. Sin embargo, debido a la gran cantidad de datos faltantes de la muestra, tuvimos que considerar otras alternativas que evitasen el ajuste de modelos con muy pocos datos. En el Cuadro B.1 del Apéndice B se recogen el número de valores perdidos de cada una de las variables explicativas. Se puede apreciar claramente que variables como X_7, X_9 o X_{13}, tienen una cantidad muy elevada de datos faltantes. Algunas de ellas suponen prácticamente la mitad de la muestra si tenemos en cuenta que contamos con un total de 672 observaciones.

Ajuste del modelo logístico 1

Dado el problema que suponían los valores perdidos de la muestra, se probó como primera alternativa la creación de un modelo logístico que contuviese a las variables X_1, X_2, X_3, X_6, X_7, X_9, X_{11}, X_{14}, X_{15}, X_{16}, X_{17}, X_{20}, X_{21}, X_{23}, X_{24} y X_{25}, pues comprobamos que todas eran estadísticamente significativas bajo modelos logísticos simples (con una única variable explicativa). Se pueden consultar los ajustes de estas regresiones simples en el Cuadro C.1. Al aplicar un procedimiento de selección de variables mediante el criterio del AIC, este primer modelo se simplificó al seleccionar 7 de las 16 variables explicativas, concretamente, X_1, X_2, X_9, X_{11}, X_{21}, X_{23} y X_{25}. Se pueden consultar los valores de los coeficientes estimados y su desviación típica, así como el estadístico de Wald y los niveles críticos de este primer modelo logístico múltiple en el Cuadro C.2.

Las variables X_{21} y X_{23} presentaban una alta correlación, observable en el Cuadro 2.10 a través de los factores de inflación, lo que nos dejaba ante un problema de colinealidad que puede afectar a la significación de las variables predictoras. Se probaron entonces dos alternativas: (1) Suprimir la variable X_{23} y aplicar un procedimiento de selección de variables como el AIC; (2) Suprimir la variable X_{21} y aplicar de nuevo el método AIC.

X_1	X_2	X_9	X_{11}	X_{21}	X_{23}	X_{25}
1.44	1.26	1.46	1.29	10.27	10.08	1.14

Cuadro 2.10: Factores de inflación de las variables explicativas del Modelo 1.

La primera alternativa nos deja ante un modelo al que llamaremos Modelo 1.A, compuesto por 3 variables explicativas: X_1, X_{21} y X_{25}. Para su ajuste se usaron un total de 271 observaciones (126 casos y 145 controles), su AIC es de 335.17, con una deviance nula de 374.35 y deviance residual de 327.17. La segunda alternativa resulta en el Modelo 1.B, con 4 covariables: X_1, X_9, X_{11} y X_{25}. Se han utilizado 87 observaciones para su diseño (42 casos y 45 controles), y tiene un AIC de 96.24, una deviance nula de 120.50 y una deviance residual de 86.24. Para comprobar su bondad de ajuste y la validez de la hipótesis nula de que el modelo seleccionado se ajusta bien a los datos, se calcularon diferentes pseudo R^2 y se aplicó el test de Hosmer-Lemeshow en ambos modelos.

En términos generales, todos los pseudo R^2 son bastante bajos (ver Cuadro 2.11), pero el Modelo 1.B presenta valores más altos en todos ellos, especialmente en Veall-Zimmermann y Nagelkerke. A mayores, el Modelo 1.A no supera el contraste de Hosmer-Lemeshow al presentar un valor del estadístico de contraste y un *p-valor* de 19.83 y 0.01, respectivamente. El Modelo 1.B, sin embargo, sí supera este

test, con un valor de 8.28 para su estadístico y un *p-valor* de 0.40.

Test	***Modelo 1.A***	***Modelo 1.B***
McFadden	0.12	0.28
McFadden ajustado	0.10	0.20
Veall-Zimmermann	0.26	0.49
Nagelkerke	0.21	0.43

Cuadro 2.11: Medidas de bondad de ajuste de los modelos 1.A y 1.B.

Ante estos resultados, escogemos el Modelo 1.B como alternativa al Modelo 1, por lo que pasará desde ahora a recibir este nombre. En el Cuadro 2.12 podemos consultar los factores de inflación de las variables explicativas que lo conforman, y observar que en este modelo no existen problemas de colinealidad. El modelo supera además las pruebas de linealidad en todas las variables continuas que lo conforman (edad y H_2O_2), apreciable en el gráfico D.1. Las distancias de Cook de la Figura D.2, nos informan de la ausencia de valores influyentes en el ajuste.

X_1	X_9	X_{11}	X_{25}
1.45	1.28	1.24	1.11

Cuadro 2.12: Factores de inflación de las variables explicativas del Modelo 1.

Las estimaciones por máxima verosimilitud de los coeficientes del modelo aparecen en el Cuadro 2.13, además de su desviación típica, los valores del estadístico de Wald y sus valores críticos. Por simplicidad, denotaremos como $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ y $\hat{\beta}_4$ a los coeficientes que acompañan a cada una de las variables explicativas X_1, X_9, X_{11} y X_{25} respectivamente. Los resultados obtenidos muestran que los coeficientes $\hat{\beta}_2$ y $\hat{\beta}_3$ no son significativamente distintos de cero.

	$\hat{\beta}$	$\hat{\sigma}$	*z value*	$Pr(>\|z\|)$
$\hat{\beta}_0$	-3.62	1.45	-2.51	0.012
$\hat{\beta}_1$	0.10	0.03	3.53	0.0004
$\hat{\beta}_2$	-0.07	0.05	-1.45	0.15
$\hat{\beta}_3$	0.99	0.60	1.63	0.10
$\hat{\beta}_4$	-0.005	0.002	-3.33	0.0009

Cuadro 2.13: Ajustes de Modelo 1.

El Cuadro 2.14 contiene las exponenciales de estos coeficientes $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ y sus intervalos de confianza, calculados mediante el método profile likelihood explicado en el capítulo anterior.

	$\widehat{OR}$	$\widehat{IC}$ *(95 %)*
$e^{\hat{\beta}_1}$	1.10	(1.05-1.17)
$e^{\hat{\beta}_2}$	0.93	(0.84-1.02)
$e^{\hat{\beta}_3}$	2.68	(0.86-9.34)
$e^{\hat{\beta}_4}$	0.10	(0.99-0.10)

Cuadro 2.14: Estimadores de los $\widehat{OR}$ y sus respectivos intervalos de confianza al 95 % para las variables del Modelo 1.

Las exponenciales de los coeficientes $\hat{\beta}$ resultantes, nos informan de que la odds de tener cáncer de mama se incrementa en 1.10 (o lo que es lo mismo, un 10 %) por cada año cumplido. El uso de anticonceptivos orales también parece ser una variable de riesgo, pues la $\widehat{OR}$ de desarrollar cáncer de

mama es 2.68 veces superior para aquellas mujeres que han tomado estos fármacos en comparación a las que nunca los han usado. Sin embargo, esta es una variable que no aparece como significativa en el Cuadro 2.13. Sucede lo mismo con X_9, meses lactando, que tampoco es significativamente distinta a cero bajo el test de Wald, pero se considera como una variable de protección al presentar una odds de 0.93. Esto significa que por cada mes no dando el pecho, la odds se desarrollar cáncer de mama aumenta en un 1.08 (un 8 %). Por último, X_{25}, variable que mide la cantidad de H_2O_2 en sangre, presenta un coeficiente para la odds de 0.1, por lo que el aumento de esta sustancia en sangre también es un factor de protección. Si observamos su inversa, vemos que el riesgo de cáncer aumenta en un 1.005 (un 0.5 %) a medida que esta sustancia disminuye en el organismo.

Los resultados de este modelo concuerdan con lo visto en el análisis descriptivo de la muestra. El número de meses lactando (X_9) presentaba medias más altas para los controles (ver Figura 2.9), y el Modelo 1 identifica que a mayor cantidad de meses dando el pecho, menor es el riesgo de cáncer de mama. De la misma manera, el incremento en sangre de H_2O_2 (X_{25}), supone un factor de protección en ambos casos (ver Figura 2.21). En la Figura 2.10 vimos que la cantidad de casos y controles que habían consumido anticonceptivos orales (X_{11}) era similar, y puede ser ésta la causa de que la variable no resulte significativa en el contraste de Wald. Sin embargo, se observaban muchos más controles que casos que nunca habían hecho uso de estos fármacos, y entre los casos, eran más numerosos las sujetos que los habían usado frente a las que no. El Modelo 1 identifica como factor de riesgo a esta variable, concordando por lo tanto, con las conclusiones extraídas del análisis descriptivo. Por último, la edad (X_1) es la única que presenta resultados contradictorios. El gráfico de cajas de la Figura 2.1 presenta una media más alta en controles que en casos, y en los histogramas de la Figura 2.2, parece que hay una mayor concentración de controles con edades más avanzadas. Esto significaría que la edad supone un factor de protección, no de riesgo cómo nos indica el Modelo 1. Estos resultados pueden deberse a la cantidad de observaciones eliminadas, pues cómo hemos visto se compone de 87 de 672. La Figura A.4 del Apéndice A.2 contiene los histogramas de la edad para casos y controles de la parte de la muestra utilizada por el Modelo 1. Vemos que en el grupo de controles hay más sujetos con menos de 50 años que en el grupo de casos. Por ello, el modelo interpreta que existe una relación directa entre el aumento de la edad y el riesgo de cáncer de mama.

Ajuste del modelo logístico 2

Otra de las opciones que manejamos fue comenzar por un modelo logístico que contuviese todas las variables de la muestra exceptuando aquellas con una cantidad muy grande de valores perdidos. Las variables X_4, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{25}, X_{26} y X_{27}, fueron suprimidas directamente como posibles predictoras al contar con 100 o más observaciones perdidas (ver Cuadro B.1 del Apéndice B). Una vez aplicada una selección de variables, solo X_1, X_5, X_6, X_{10}, X_{17}, X_{19} y X_{28} pasaron a formar parte de este segundo modelo logístico al que llamaremos Modelo 2.

El Modelo 2 presenta un AIC de 762.61, y una deviance nula y residual de 816.98 y 746.61, respectivamente. Los resultados del Cuadro 2.15 reflejan una baja correlación entre las variables explicativas del modelo, lo que indica ausencia de colinealidad. El gráfico (a) de la Figura D.4 demuestra la ausencia de valores atípicos en la muestra utilizada. Sin embargo, la suposición de linealidad se ve violada por las variables X_{10} y X_{19}, IMC y monocitos bajo medidas porcentuales. En los gráficos de dispersión de la Figura D.3 podemos ver claramente cómo la relación establecida entre las variables X_1 y X_{17} (edad y neutrófilos en valor porcentual, respectivamente) y la transformación logarítmica de las probabilidades, es lineal. Esta relación no está tan clara para X_{10} y X_{19}. A pesar de esto, la ausencia de colinealidad y de valores atípicos hace que nos decantemos por considerar el modelo como válido para el análisis.

X_1	X_5	X_6	X_{10}	X_{17}	X_{19}	X_{28}
1.33	1.75	1.82	1.08	1.13	1.10	1.07

Cuadro 2.15: Factores de inflación de las variables explicativas del Modelo 2.

Su bondad de ajuste es menor que la del Modelo 1, pues como podemos ver en el Cuadro 2.16, los pseudo R^2 son más bajos. Sin embargo, el *p-valor* del contraste de Hosmer-Lemeshow es mayor, siendo este de 0.58 frente a 0.40 del Modelo 1.

Test	***Modelo 2***
McFadden	0.09
McFadden ajustado	0.07
Veall-Zimmermann	0.18
Nagelkerke	0.15

Cuadro 2.16: Medidas de bondad de ajuste del Modelo 2.

En cuanto al valor explicativo de las variables, los resultados del Cuadro 2.17 reflejan que todas ellas son significativamente distintas a cero. En este caso, los coeficientes $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$, $\hat{\beta}_7$, se corresponden con las variables X_1, X_5, X_6, X_{10}, X_{17}, X_{19} y X_{28}, de manera respectiva.

	$\hat{\beta}$	$\hat{\sigma}$	*z value*	$Pr(>\|z\|)$
$\hat{\beta}_0$	-3.76	1.04	-3.62	0.0003
$\hat{\beta}_1$	-0.03	0.008	-4.28	0.00002
$\hat{\beta}_2$	0.58	0.32	1.78	0.07
$\hat{\beta}_3$	-0.20	0.09	-2.14	0.03
$\hat{\beta}_4$	0.03	0.02	2.06	0.04
$\hat{\beta}_5$	0.05	0.01	4.64	0.000004
$\hat{\beta}_6$	0.14	0.05	2.56	0.01

$\hat{\beta}_7$	0.005	0.002	2.14	0.03

Cuadro 2.17: Ajustes del Modelo 2.

El Cuadro 2.18 contiene las exponenciales de los coeficientes $\hat{\beta}$ arriba mostrados, así como sus intervalos de confianza calculados mediante el método profile likelihood.

	$\widehat{OR}$	$\widehat{IC}$ *(95 %)*
$e^{\hat{\beta}_1}$	0.97	(0.95-0.98)
$e^{\hat{\beta}_2}$	1.78	(0.95-3.38)
$e^{\hat{\beta}_3}$	0.82	(0.67-0.98)
$e^{\hat{\beta}_4}$	1.03	(1.002-1.07)
$e^{\hat{\beta}_5}$	1.05	(1.03-1.07)
$e^{\hat{\beta}_6}$	1.15	(1.03-1.28)
$e^{\hat{\beta}_7}$	1.005	(1.0004-1.01)

Cuadro 2.18: Estimadores de los $\widehat{OR}$ y sus respectivos intervalos de confianza al 95 % para las variables del Modelo 2.

Como vemos, la odds asociada a la variable X_1, edad al diagnóstico o entrevista, es de 0.97, lo que significa que cada año cumplido disminuye el riesgo de sufrir cáncer de mama. Al calcular su inversa, vemos que la odds de la edad tiene un valor de 1.03, por lo que esta aumenta en un 3 % por cada año menos. El coeficiente $\hat{\beta}_2$ determina que la odds asociada a la maternidad aumenta en un 1.78 (un 78 %) el riesgo de cáncer de mama en comparación a mujeres que no han tenido hijos. Sin embargo, este riesgo disminuye cuantos más partos se hayan tenido, pues la variable X_6, número de hijos, presenta una odds de 0.82. Esto significaría que, para cualquiera de las mujeres de la muestra, el riesgo de cáncer de mama aumenta en un 1.22 (un 22 %) por cada hijo no nacido. Por otro lado, las variables X_{10}, X_{17} y X_{28}, son consideradas de riesgo. Si observamos los coeficientes de cada una de ellas, veremos que cada unidad de IMC (X_{10}), aumenta en un 1.03 (o el 3 %) la odds de cáncer de mama, al igual que el colesterol (X_{28}), con una odds de 1.005 (un 0.5 %). Sin embargo, estos valores están muy cercanos a 1, valor que indica falta de asociación entre las variables explicativas y la respuesta. Sucede lo mismo con el aumento porcentual de neutrófilos en sangre (variable X_{17}), que aumenta el riesgo de cáncer de mama en un 1.05 (un 5 %). Los monocitos, por otro lado (variable X_{19}) presentan una odds más alta, aumentando en un 1.15 (o un 15 %) el riesgo de la enfermedad.

Estos resultados coinciden con lo visto en el análisis exploratorio. Se aprecia en la Figura 2.2, que la cantidad de casos diagnosticados a partir de los 70 años es mucho menor en comparación a edades previas. Por lo que a partir de esta edad, el riesgo de cáncer disminuye. Si recordamos, en el ajuste del Modelo 1, la odds asociada a esta variable aumentaba a medida que se cumplían años, mientras que en este caso sucede lo contrario. Esto posiblemente sea debido a las diferencias entre las observaciones que manejan ambos modelos como hemos visto en la subsección anterior.

En cuanto a la maternidad y al número de hijos, los resultados del Modelo 2 concuerdan con las conclusiones del análisis descriptivo de la muestra, donde el Cuadro 2.4 y la Figura 2.7, anunciaban al número de hijos como variable de protección, con una media superior en el grupo control.

El IMC y el colesterol no parecían presentar diferencias muy notables entre casos y controles en el análisis gráfico. Se apreciaba en las Figuras 2.13 y 2.22 que ambas variables son similares en distribución y en media para ambos grupos (ver también Cuadros 2.5 y 2.8). El Modelo 2 sitúa estas variables como factores de riesgo, pero como hemos visto, sus odds son cercanas a 1 (1.03 y 1.005, respectivamente). Ante estos valores, concluimos que ambos métodos de análisis aportan los mismos resultados, pues la regresión arroja una relación positiva débil entre estas variables y el riesgo el cáncer de mama. La proporción de neutrófilos en sangre sin embargo, sí suponía un factor de riesgo muy claro en el análisis descriptivo (ver Figura A.1 y Cuadro 2.7). Las conclusiones del análisis de regresión se inclinan también en esta dirección, pero de nuevo la odds no es muy alta (1.05). En cuanto a la proporción de monocitos en sangre, los resultados de la regresión no concuerdan del todo con el análisis descriptivo de la variable. Hemos visto que tanto en media como en distribución, las diferencias no son muy acusadas

entre ambos grupos (ver Cuadro 2.7 y Figura A.2), por lo que a priori esta variable no tendría una influencia clara sobre el riesgo de cáncer de mama. Sin embargo, el Modelo 2 contradice estos resultados al establecer una odds con un valor de 1.15 para esta variable, lo que indicaría que cantidades altas de monocitos en sangre supodrían un factor de riesgo del cáncer de mama.

2.3. Predicciones y curvas ROC

Para evaluar la capacidad predictiva de los modelos 1 y 2 ajustados, estudiaremos las curvas ROC y el AUC asociados. Como punto de corte c para la clasificación, hemos maximizado el índice de Youden. Debemos recordar que para este trabajo hemos utilizado las mismas observaciones para generar los modelos de regresión logística 1 y 2, y sus correspondientes curvas ROC y AUC. Esto provoca que los resultados obtenidos por el AUC sean demasiados optimistas y no nos permitan valorar la capacidad predictiva real de cada modelo. Por ello, hemos utilizado la validación cruzada para corregir los valores del AUC, dejando una observación fuera en cada ejecución del algoritmo, tal y como se explica en la Sección 1.4. Para efectuar el cálculo de las curvas ROC y el AUC aparente, se han utilizado los paquetes `ROCR` y `pROC` de RStudio. Este último ha resultado muy útil para obtener el c que maximiza el índice de Youden en cada caso, y para el cálculo de la sensibilidad, la especificidad y la precisión de cada modelo.

Modelo 1

La curva ROC y el AUC aparente (que utiliza todas las observaciones de la muestra) del Modelo 1, con las variables predictivas X_1, X_9, X_{11} y X_{25}, presenta un punto de corte óptimo $c = 0.60$, una sensibilidad de 0.71 y una especificidad del 0.89, lo que sitúa su precisión en un 0.80. Su índice de Youden es igual a 0.60. Sin embargo, al corregir la curva ROC utilizando la validación cruzada, los valores anteriores sufren modificaciones. El punto o umbral óptimo pasa a ser $c = 0.57$, colocando la sensibilidad y la especificidad del modelo en 0.71 y 0.8 respectivamente. Esto nos deja con un índice de Youden igual a 0.51 y una precisión de 0.76, ambos valores más bajos que los reportados bajo el ajuste convencional.

La Figura 2.24 contiene las dos curvas ROC anteriores, con los puntos c óptimos que maximizan el valor del índice de Youden en cada caso. Al ser ambas diferentes, el valor de su área bajo la curva es también distinta. La curva ROC aparente presenta un AUC = 0.85, con intervalos de confianza (0.77-0.93), mientras la corregida tiene un valor de 0.79, con los intervalos de confianza (0.7-0.89).

En el Cuadro 2.19 podemos ver la matriz de confusión que surge una vez corregido el AUC. En esta matriz apreciamos la cantidad de aciertos y fallos que comete nuestro modelo binario al clasificar, de la forma que hemos presentado en el Cuadro 1.4. Como hemos mencionado con anterioridad, los casos

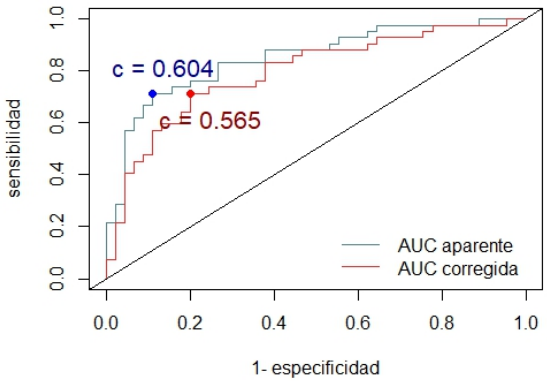


Figura 2.24: Curvas ROC y AUC aparente y corregido del Modelo 1.

se corresponden con mujeres enfermas y los controles con mujeres sanas. De este modo, las variables diagnóstico escogidas clasifican bien a 66 de las 87 observaciones utilizadas por el Modelo 1, esto es, aproximadamente a un 76 % de las sujeto de la muestra, coincidiendo este valor con la precisión. Dado que la cantidad de controles bien clasificados es superior a la de casos, la especificidad es más alta tal y como hemos visto al principio de este apartado (0.8 frente a 0.71, respectivamente).

	Estado real	
Diagnóstico	**Controles**	**Casos**
Controles	36	12
Casos	9	30

Cuadro 2.19: Matriz de confusión de casos y controles reales frente a casos y controles diagnosticados por el Modelo 1.

En la Figura 2.25 aparece representada esta matriz de confusión mediante un gráfico de cajas (izquierda) y un gráfico de dispersión (derecha). En el diagrama de cajas se enfrentan las predicciones del modelo para cada observación (eje y) con la condición de caso o control (eje x) de cada una de ellas. La línea roja horizontal se corresponde al punto de corte $c = 0.57$. Las observaciones por encima de esta línea son consideradas casos, y por debajo, controles. Los controles situados por encima del punto c son falsos positivos. Los casos situados por debajo de c son falsos negativos. En base a las predicciones del modelo, la mayoría de las observaciones son clasificadas correctamente en sus respectivos grupos, identificando a gran parte de los casos como sujetos enfermos, y a controles como sujetos sanos. En el gráfico de dispersión vuelven a enfrentarse las predicciones del modelo (eje y) frente a cada una de las observaciones (eje x), representando su grupo de pertenencia en base al color. Las cruces rojas se corresponden con los casos, y las azules con los controles. La franja roja dibujada se corresponde con el punto de corte $c = 0.57$ que hemos fijado para el cálculo de la especificidad y la sensibilidad. Las observaciones situadas por encima de c son consideradas casos, y las situadas por debajo, controles. Por lo tanto, las cruces rojas que se encuentran por debajo de c son falsos negativos, y de la misma manera, las cruces azules por encima de c, son falsos positivos.

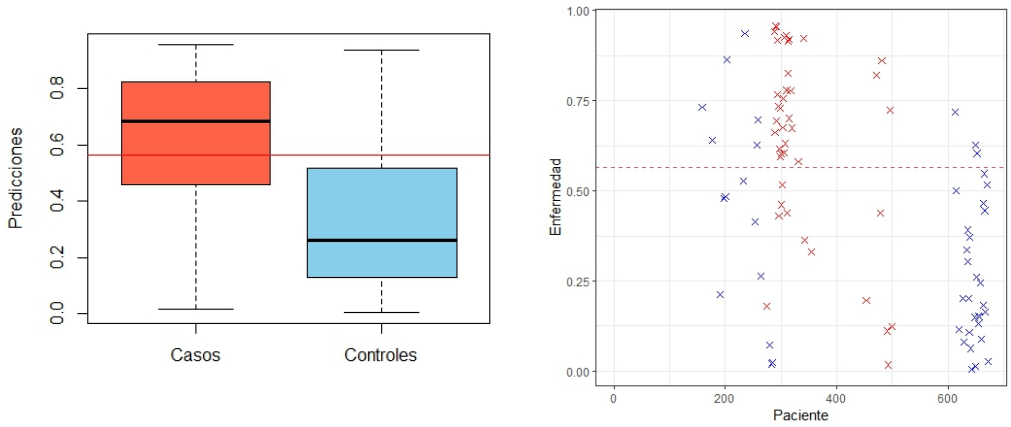


Figura 2.25: Diagrama de cajas (izquierda) y gráfico de dispersión de la matriz de confusión (derecha) con $c = 0.57$, ambas representaciones de la especificidad y sensibilidad del Modelo 1.

Modelo 2

Bajo los cálculos del AUC aparente, el Modelo 2, compuesto por las variables X_1, X_5, X_6, X_{10}, X_{17}, X_{19} y X_{28}, posiciona su c óptimo en 0.46. Su sensibilidad es igual a 0.68, y su especificidad 0.63, con una precisión de 0.65. Su índice de Youden es igual a 0.31, un valor no muy alto e inferior al del Modelo 1 bajo el AUC aparente. Al realizar las correcciones pertinentes al cálculo de la curva ROC y al AUC utilizando la validación cruzada dejando una observación fuera, el umbral óptimo pasa a ser $c = 0.46$, por lo que la sensibilidad y la especificidad de este sistema de clasificación tienen los valores 0.66 y 0.62, respectivamente. Esto modifica a su vez el índice de Youden, que ahora es igual a 0.28. Su precisión es, por lo tanto, de 0.64.

En la Figura 2.26 vemos representadas ambas curvas ROC con sus respectivos c óptimos, también calculados siendo el valor que maximiza al índice de Youden. El AUC aparente tiene un valor de 0.69, mientras que el corregido de 0.67. Sus intervalos de confianza se sitúan, para el AUC aparente entre (0.65-0.74), y para el corregido, (0.63-0.72). La longitud de los intervalos de confianza es menor que los calculados para el Modelo 1.

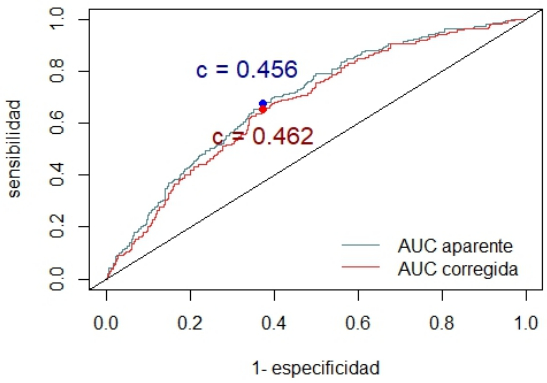


Figura 2.26: Curvas ROC y AUC aparente y corregido del Modelo 2.

El Cuadro 2.20 se corresponde con la matriz de confusión de este Modelo 2. Vemos que comete más fallos de clasificación que el modelo anterior, perceptible además en los gráficos de la Figura 2.27, posiblemente debido a que la cantidad de observaciones es mucho mayor en comparación. Se aprecia en ambas figuras que la mayor parte de las observaciones se clasifican en el grupo correcto, sin embargo la cantidad de falsos positivos y negativos es muy alta. Si nos fijamos en el gráfico de cajas, la distribución de las probabilidades de cáncer de casos y controles es muy similar, lo que dificulta la correcta clasificación de las observaciones. De hecho, hemos podido comprobar al inicio que los niveles de sensibilidad y especificidad son bastante bajos (aproximadamente de 0.66 y 0.63,

respectivamente), lo que nos deja ante una proporción de falsos negativos y positivos bastante alta (0.34 y 0.37, respectivamente). La precisión de este modelo es de hecho, bastante baja, de solo 0.64, lo que significa que el porcentaje de fallo es de casi el 36 %.

	Estado real	
Diagnóstico	**Controles**	**Casos**
Controles	197	95
Casos	117	182

Cuadro 2.20: Matriz de confusión de los casos y los controles reales frente a los casos y los controles predichos por el Modelo 2.

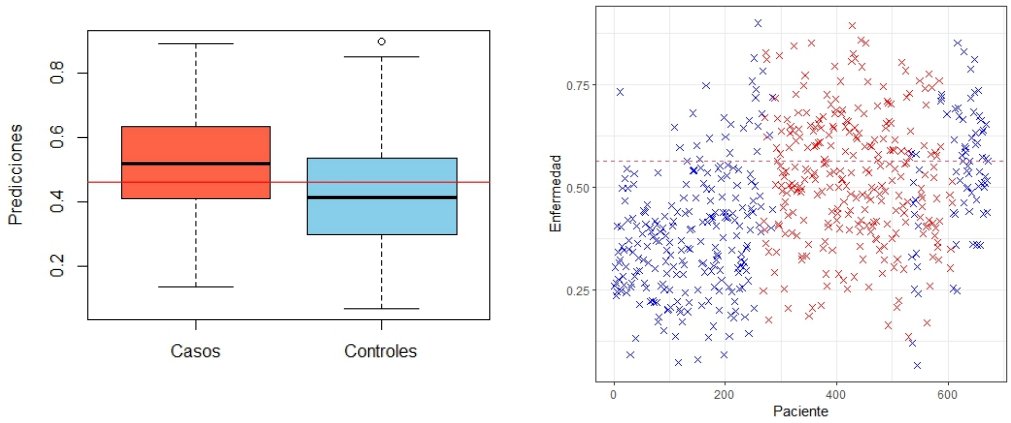


Figura 2.27: Diagrama de cajas (izquierda) y gráfico de dispersión de la matriz de confusión (derecha) con $c = 0.46$, ambas representaciones de la especificidad y sensibilidad del Modelo 2.

En definitiva, los modelos de regresión presentados en esta sección indican que el uso de anticonceptivos orales, la maternidad y un aumento en el porcentaje de monocitos en sangre, son factores de riesgo. Un alto porcentaje de neutrófilos y niveles de colesterol e IMC altos también aumentan ligeramente el riesgo de cáncer. El incremento en el número de hijos y los meses dando el pecho, sin embargo, suponen factores de protección, al igual que el H_2O_2, que disminuye el riesgo de cáncer de mama a medida que sus niveles aumentan en sangre. El impacto de la edad no está claro. El Modelo 1 clasifica esta variable de riesgo, mientras el Modelo 2 como factor de protección. En cuanto a los modelos propuestos, ambos presentan propiedades positivas y negativas como modelos de regresión logística para el análisis epidemiológico de la base de datos de este trabajo. El Modelo 1 cuenta con una mayor precisión como clasificador entre casos y controles, y un AUC superior al del Modelo 2. En el Cuadro 2.21 podemos ver a modo de comparación, la cantidad de observaciones (n), el c, y los índices de sensibilidad, especificidad, precisión, Youden y AUC de ambos modelos utilizando la corrección de la curva ROC mediante validación cruzada.

	n	***c***	**sensibilidad**	**especificidad**	**precisión**	**Youden**	**AUC**
Modelo 1	87	0.56	0.71	0.8	0.76	0.51	0.79
Modelo 2	591	0.46	0.66	0.63	0.64	0.28	0.67

Cuadro 2.21: Tabla comparativa entre el número de observaciones (n), punto de corte (c) y los índices de sensibilidad, especificidad, precisión, Youden y AUC de los Modelos 1 y 2.

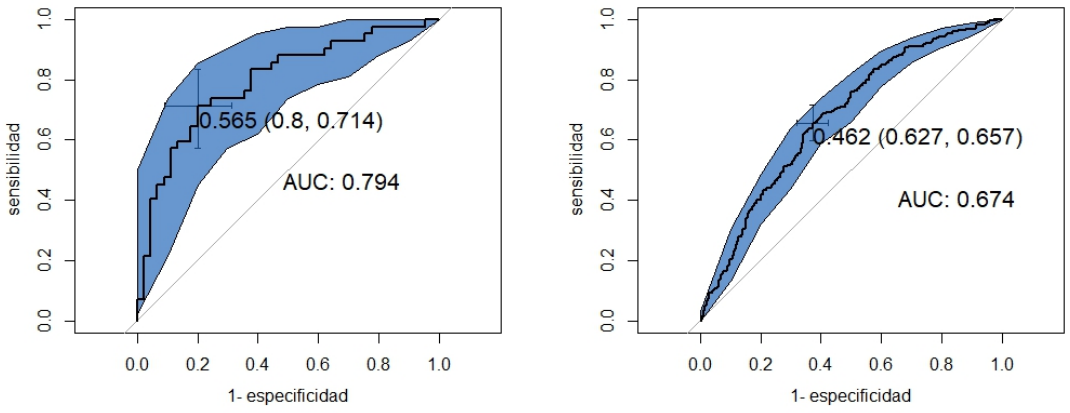
Además, el AIC del Modelo 1 es más bajo que el del Modelo 2 (96.244 y 762.61, respectivamente) y presenta una deviance nula y residual también más bajas. En términos de bondad de ajuste, el Modelo 1 presenta valores más altos en los pseudo R^2 calculados. Sin embargo, Smith y Kenna (2013) ya anunciaban en su estudio que estos pseudo R^2 pueden ser susceptibles a cambios en el tamaño muestral, por lo que en nuestro caso, podrían no ser muy útiles para comparar ambos modelos. En el Cuadro 2.22 se pueden observar los valores de estos test.

Bajo los resultados de los test implementados y el análisis de las curvas ROC de ambos modelos, parece que el Modelo 1 presenta mejores propiedades como método de predicción de casos nuevos de cáncer de mama. Sin embargo, este sufre de un defecto que podría ser considerado muy grave, y es que la cantidad de observaciones que suprime de la muestra original es muy alta. Vemos en el Cuadro 2.21, que mientras el Modelo 2 solo suprime 81 observaciones de la base de datos original, el Modelo 1 lo

Test	***Modelo 1***	***Modelo 2***
McFadden	0.28	0.09
McFadden ajustado	0.20	0.07
Veall-Zimmermann	0.49	0.18
Nagelkerke	0.43	0.15

Cuadro 2.22: Bondad de ajuste de los modelos 1 y 2.

hace con 585, lo que puede generar resultados sesgados y poco representativos de la muestra de partida y por consiguiente, de la población de estudio. El Modelo 2, a pesar de no contar con propiedades de clasificación y bondad de ajustes tan altas, se mantiene más fiel a la realidad de la base de datos y refleja con más realismo la complejidad que supone el diagnóstico del cáncer de mama. La Figura 2.28 nos muestra los intervalos de confianza de las curvas ROC y el AUC corregido de cada uno de los modelos.



(a) Modelo 1 (b) Modelo 2

Figura 2.28: Intervalos de confianza de las curvas ROC de los modelos 1 y 2.

Si observamos detenidamente, veremos que la banda de confianza del Modelo 1 es mucho más amplia que la del Modelo 2. La diferencia en la cantidad de datos utilizados en ambos modelos provoca

que el Modelo 2 sea más preciso a la hora de estimar sus posibles curvas ROC en comparación al Modelo 1.

Capítulo 3

Conclusiones

En este trabajo hemos podido comprobar la dificultad que supone el estudio de fenómenos tan complejos como el cáncer. Ya desde un principio, la literatura epidemiológica consultada anunciaba la dificultad que supone el establecer relaciones entre determinados factores de exposición y el riesgo de desarrollar esta enfermedad. Y es que hasta la fecha, el cáncer de mama puede ser considerada una enfermedad multicausal, no estando todavía del todo claros los factores que aumentan o disminuyen las probabilidades de su aparición. Las investigaciones sobre este tema se contradicen en el impacto que tienen muchas de las variables que hemos visto. En algunos estudios aparecen como factores de riesgo y en otros, no presentan relevancia alguna. El análisis realizado en este trabajo cuerda con algunos de ellos, destacando como factores de riesgo variables como el NLR, el PLR, una proporción alta de neutrófilos en sangre, el tener antecedentes familiares de cáncer de mama y/o ovárico, el uso de anticonceptivos orales, la maternidad, el retraso en la edad al primer hijo nacido y una menarquia anterior a los 13 años. Como factores de protección encontramos la edad, la menopausia, el número de hijos, la lactancia y niveles altos de linfocitos y de H_2O_2. Variables cómo el IMC, la altura, el peso, la terapia hormonal sustitutiva, el alcohol, el tabaco, el colesterol, los monocitos y las plaquetas, no presentan una relación clara con el cáncer de mama a pesar de que muchas de ellas suponen factores de riesgo desde la literatura epidemiológica consultada.

Debemos mencionar que al inicio de este trabajo, se pretendía hacer un especial énfasis en el estudio de la altura como factor de riesgo del cáncer de mama. Sin embargo, nuestros análisis no concordaban con la literatura epidemiológica disponible sobre esta variable, donde resultaba tener alta relevancia en el riesgo de este tipo de cáncer. Algunas de estas investigaciones estudiaban el efecto de esta variable junto con otras como el peso, el IMC y el estado menopáusico, que podrían estar ejerciendo influencia sobre los resultados obtenidos. Por ello, tras realizar numerosos análisis descriptivos y de regresión, hemos enfocado este trabajo desde un punto de vista más generalista, dando el mismo protagonismo

a cada variable de la muestra de estudio.

A mayores de la complejidad que supone el estudio del cáncer de mama, la base de datos utilizada en este trabajo contiene mucha información faltante, lo que ha dificultado todavía más el análisis de los fenómenos que pueden influir en la aparición de la enfermedad, y llevarnos a conclusiones sesgadas y poco representativas de la población de estudio. La cantidad de valores perdidos ha supuesto un problema desde el inicio, obligándonos a buscar alternativas para diseñar una regresión logística múltiple que, sin poder evitarlo, deja atrás mucha información que podría ser de interés. En el Modelo 1, construido en su inicio por variables significativas bajo modelos de regresión logística simples, un total de 585 observaciones han tenido que ser suprimidas para poder establecer un modelo que tuviese en cuenta la mayor cantidad de variables posibles. En el Modelo 2, se han ignorado directamente un total de 12 covariables de las 28 originales, lo que supone de nuevo una pérdida de información muy elevada. Esto supone un riesgo de cara a la bondad de ajuste de los modelos de regresión, pues sus resultados pueden estar sesgados y llevarnos a conclusiones incorrectas. Además, muchas de las variables no han podido ser analizadas mediante la regresión logística, teniendo que sacar conclusiones únicamente del análisis descriptivo. Con bases de datos como la utilizada en este trabajo, se vuelve necesario el uso de métodos o procedimientos con los que tratar cantidades grandes de datos faltantes. Rubin (1987) presenta la técnica estadística de la imputación múltiple, diseñada para aprovechar la flexibilidad de la computación moderna para generar los datos que faltan. A raíz de esta, han surgido dos enfoques generales para la imputación de datos multivariantes: el modelado conjunto y la imputación múltiple por ecuaciones encadenadas (Buuren, 2011).

A pesar de todo esto, no debemos despreciar la capacidad predictiva del modelo logístico, que bajo condiciones más favorables, es muy útil en la epidemiología tal y como hemos visto en el Capítulo 1. De los dos modelos finalmente establecidos, el Modelo 1 presenta una mayor precisión para clasificar casos y controles. Hemos visto en el estudio de su curva ROC y su AUC, que es un modelo con capacidades predictivas notables, pero que ha sido ajustado con pocos datos. Por otro lado, el Modelo 2 es peor a la hora de predecir y clasificar casos y controles, pero utiliza una muestra más representativa de la realidad del cáncer, que aunque no cuenta con variables que podrían ser decisivas, la cantidad de valores perdidos es muy baja. Este modelo podría ajustarse mejor a la base de datos estudiada y quizás reflejar con mayor precisión la complejidad que supone el diagnóstico de casos de cáncer de mama.

Apéndice A

Análisis descriptivo adicional

A.1. Variables medidas bajo niveles porcentuales

En este apéndice incluimos los análisis gráficos de las variables X_{17}, X_{19} y X_{21}, que miden la proporción de neutrófilos, monocitos y linfocitos en $10^3/\mu_L$ de sangre, respectivamente.

En la Figura A.1, la proporción de neutrófilos es mayor en el grupo casos que en el de controles, lo que significa que los primeros tienen un nivel más alto de neutrófilos en sangre en comparación al resto de glóbulos blancos estudiados. La proporción de monocitos de la Figura A.2 no muestra diferencias muy grandes entre ambos grupos, aunque hay más casos con niveles altos de estas células. Por último, los controles tienen una media más alta y una mayor cantidad de observaciones en los cuantiles más altos de la distribución de densidad correspondiente a la Figura A.3, que representa la proporción de linfocitos de la muestra.

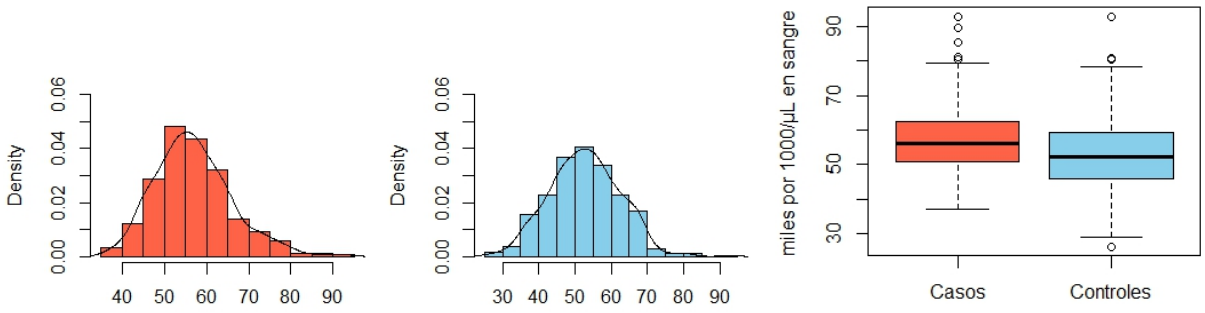


Figura A.1: Histograma de la proporción de neutrófilos en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

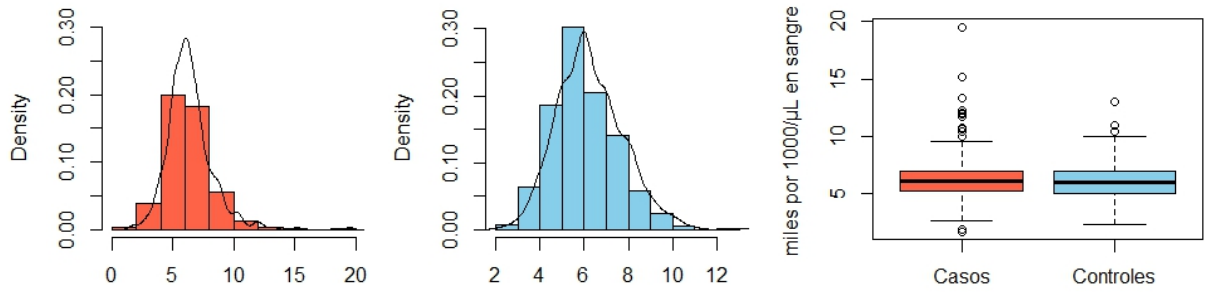


Figura A.2: Histograma de la proporción de monocitos en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

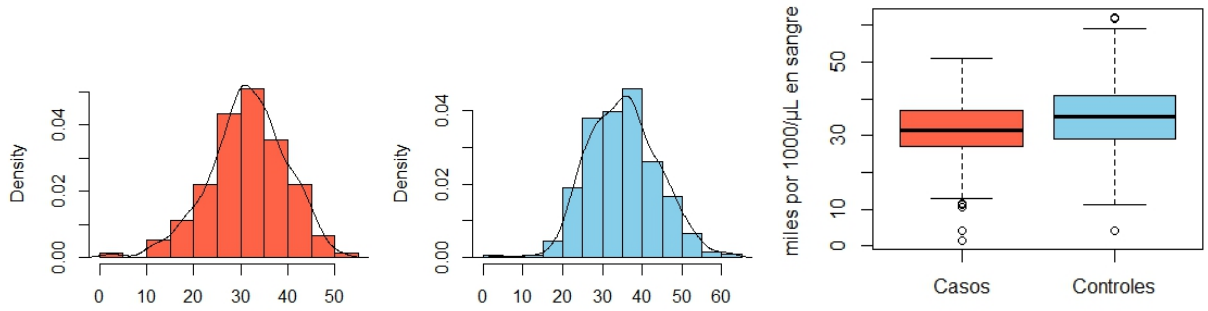


Figura A.3: Histograma de la proporción de linfocitos en $10^3/\mu_L$ de sangre para casos (izquierda), controles (centro), y gráfico de cajas para la misma variable (derecha).

A.2. Edad al diagnóstico/entrevista del Modelo 1

En esta última Figura A.4, se encuentran los histogramas de casos y controles para la variable edad al diagnóstico/entrevista que forman parte de las observaciones utilizadas por el Modelo 1.

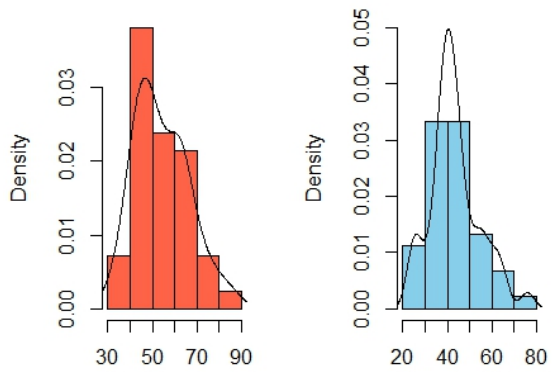


Figura A.4: Histograma de la edad al diagnóstico para casos (izquierda) y controles (derecha) de las observaciones incluidas en el Modelo 1.

Apéndice B

Valores perdidos

En el Cuadro B.1 se pueden apreciar la cantidad de valores faltantes de cada una de las variables tenidas en cuenta para el análisis. X_i se corresponde con cada una de las variables de la muestra, y *NA's* el número de observaciones faltantes de cada una de ellas.

X_i	*NA's*	X_i	*NA's*	X_i	*NA's*
X_1	0	X_{11}	232	X_{21}	14
X_2	62	X_{12}	233	X_{22}	13
X_3	0	X_{13}	250	X_{23}	22
X_4	297	X_{14}	303	X_{24}	16
X_5	22	X_{15}	272	X_{25}	396
X_6	23	X_{16}	14	X_{26}	100
X_7	295	X_{17}	12	X_{27}	101
X_8	281	X_{18}	15	X_{28}	35
X_9	300	X_{19}	12		
X_{10}	32	X_{20}	16		

Cuadro B.1: Recuento total de valores perdidos en cada variable de la muestra.

Apéndice C

Ajustes de los modelos de regresión logística secundarios

C.1. Regresiones logísticas simples

En el Cuadro C.1 se muestran los ajustes de los modelos logísticos simples cuyas covariables resultaron significativas, y que pasaron a formar parte del primer modelo logístico múltiple antes de aplicar el AIC.

Variable	$\hat{\beta}$	$\hat{\sigma}$	*z value*	$Pr(>\|z\|)$
X_1	-0.03046	0.00599	-5.085	0.000000367
X_2	-0.10678	0.05085	-2.100	0.0357
X_3	-0.8114	0.1707	-4.753	0.000002
X_6	-0.23051	0.06308	-3.654	0.000258
X_7	0.05703	0.02267	2.516	0.0119
X_9	-0.04518	0.01706	-2.649	0.00808

X_{11}	1.1678	0.2136	5.467	0.0000000458
X_{14}				
ex fumadora	1.1624	0.3071	3.786	0.000153
fumadora	1.2716	0.3483	3.651	0.000262
X_{15}	1.1798	0.2191	5.385	0.0000000723390297
X_{16}	0.2110	0.0577	3.657	0.000256
X_{17}	0.045717	0.008657	5.281	0.0000001283
X_{20}	-0.4260	0.1188	-3.584	0.000338
X_{21}	-0.053847	0.009752	-5.522	0.0000000335
X_{23}	0.5072	0.1061	4.781	0.000001745
X_{24}	0.006017	0.001764	3.410	0.000649
X_{25}	-0.0023142	0.0005438	-4.255	0.0000209

Cuadro C.1: Estimadores de los coeficientes $\hat{\beta}$, la desviación típica, el estadístico de Wald y los valores críticos de las variables explicativas significativas al 0.05 bajo modelos de regresión logística simple.

C.2. Regresión logística múltiple

En este Cuadro C.2 se muestran los ajustes del primer modelo logístico múltiple, construido a partir de las variables que superaron el contraste de Wald en un modelo de regresión logística simple, y que una vez aplicado el AIC, se mantuvieron en el modelo. Se observa que algunas de ellas no son

significativas, como es el caso de X_9 y X_{11}, meses lactando y anticonceptivos orales respectivamente.

Variable	$\hat{\beta}$	$\hat{\sigma}$	*z value*	$Pr(>\|z\|)$
X_1	0.1	0.03	3.07	0.002
X_2	0.42	0.25	1.66	0.1
X_9	-0.09	0.06	-1.50	0.13
X_{11}	1.15	0.71	1.63	0.10
X_{21}	-0.30	0.13	-2.32	0.02
X_{23}	-2.24	1.18	-1.91	0.06
X_{25}	-0.006	0.002	-3.25	0.001

Cuadro C.2: Estimadores de los coeficientes $\hat{\beta}$, la desviación típica, el estadístico de Wald y los valores críticos de las variables explicativas del primer modelo logístico múltiple ajustado.

C.3. Ajustes del Modelo 1.A

En el Cuadro C.3, presentamos los ajustes de los coeficientes $\hat{\beta}$ del Modelo 1.A, descartado como posible modelo de regresión logística para el análisis de la muestra al no superar el test de Hosmer-Lemeshow y presentar peores resultados que el Modelo 1.B en los pseudo R^2 utilizados.

Variable	$\hat{\beta}$	$\hat{\sigma}$	*z value*	$Pr(>\|z\|)$
X_1	0.04	0.01	3.82	0.0001
X_{21}	-0.05	0.016	-3.22	0.001

X_{25}	-0.002	0.0006	-4.26	2.02e-05

Cuadro C.3: Estimadores de los coeficientes $\hat{\beta}$, la desviación típica, el estadístico de Wald y los valores críticos de las variables explicativas del Modelo 1.A.

Apéndice D

Pruebas de linealidad y distancias de Cook

A continuación se muestran los gráficos utilizados para el análisis de la linealidad y el estudio de datos influyentes para los modelos 1 y 2. Los gráficos de linealidad se han construido enfrentando las observaciones de las variables continuas con las predicciones de cada modelo. Para las distancias de Cook se ha utilizado el código *cooks.distance* de RStudio, y se han generado otros gráficos de acompañamiento que muestran las 3 observaciones con las distancias más largas.

D.1. Modelo 1

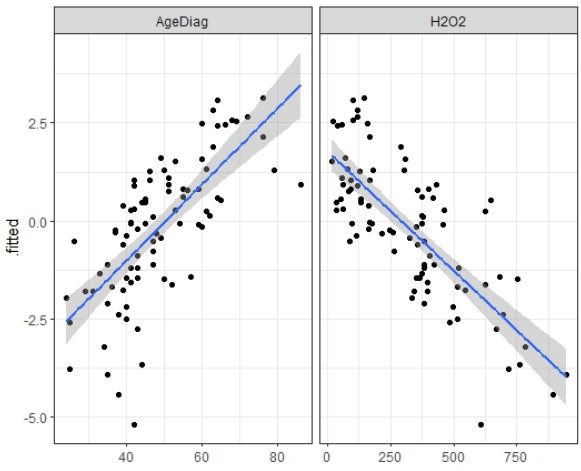
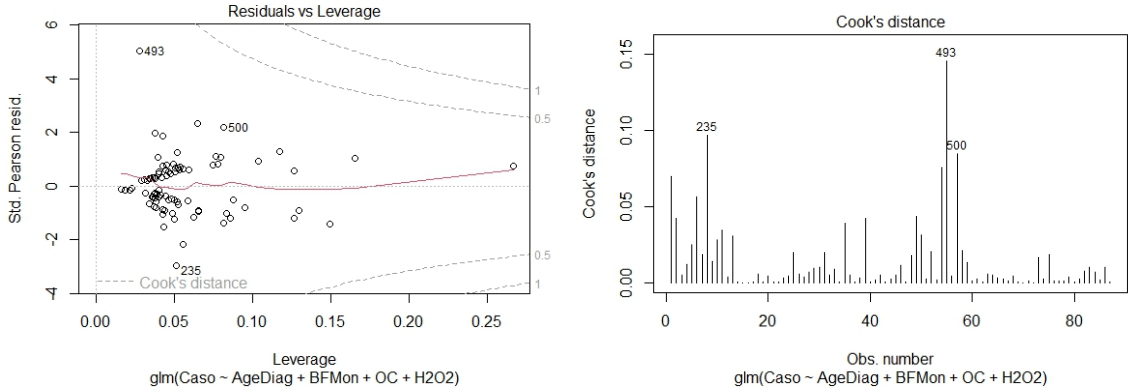


Figura D.1: Prueba de linealidad del Modelo 1.



(a) Distancias de Cook

(b) Distancias más largas

Figura D.2: Distancias de Cook del Modelo 1.

D.2. Modelo 2

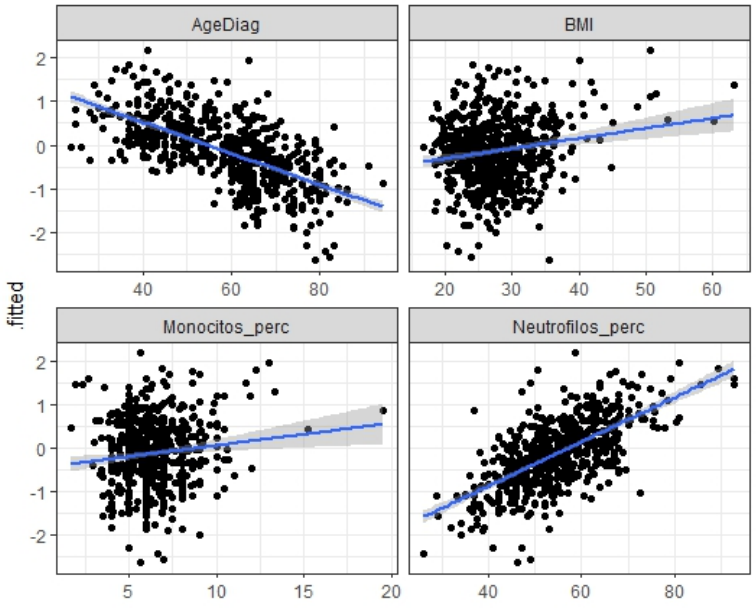
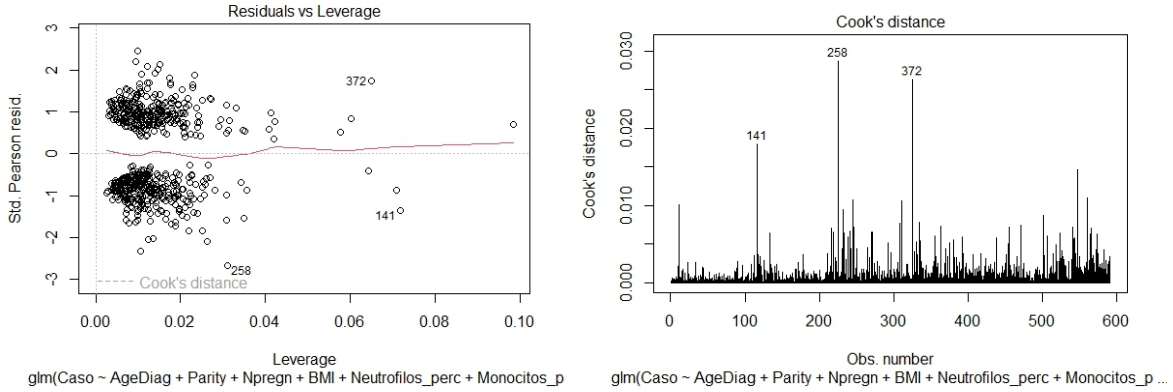


Figura D.3: Prueba de linealidad del Modelo 2.



(a) Distancias de Cook

(b) Distancias más largas

Figura D.4: Distancias de Cook del Modelo 2.

Apéndice E

Código R para la lectura de datos y su análisis

En esta apéndice se presenta el código utilizado en RStudio para el análisis de la muestra, incluyendo los paquetes utilizados, las modificaciones pertinentes hechas a la base de datos, los gráficos diseñados, las pruebas realizadas y los modelos logísticos diseñados. Las variables utilizadas para el estudio de casos-controles son las siguientes:

- Caso(Y): casos y controles
- AgeDiag (X_1): eda al diagnóstico (casos) o entrevista (controles)
- Agemenarche (X_2): edad a la menarquia
- MenoStatus (X_3): estado menopáusico (pre o post)
- AgeMenopause (X_4): edad a la menopausia
- Parity (X_5): maternidad (si o no)
- Npregn (X_6): número de hijos
- AgeFFTP (X_7): edad al primer hijo
- BF (X_8): lactancia (si o no)
- BFMon (X_9): meses lactando
- BMI (X_{10}): IMC
- OC (X_{11}): anticonceptivos orales (si o no)

- HR (X_{12}): terapia hormonal sustitutiva (si o no)
- AlcoholFreq (X_{13}): consumo de alcohol
- Smoking (X_{14}): consumo de tabaco (nunca, ex, fuma)
- FHPS (X_{15}): historial familiar de cáncer de mama y/o ovárico(si o no)
- Neutrofilos_abs (X_{16}): recuento absoluto de neutrófilos en $10^3/\mu_L$ de sangre
- Neutrofilos_perc (X_{17}): recuento porcentual de neutrófilos en $10^3/\mu_L$ de sangre
- Monocitos_abs (X_{18}): recuento absoluto de monocitos en $10^3/\mu_L$ de sangre
- Monocitos_perc (X_{19}): recuento porcentual de monocitos en $10^3/\mu_L$ de sangre
- Linfocitos_abs (X_{20}): recuento absoluto de linfocitos en $10^3/\mu_L$ de sangre
- Linfocitos_perc (X_{21}): recuento porcentual de linfocitos en $10^3/\mu_L$ de sangre
- Plaquetas_abs (X_{22}): recuento absoluto de plaquetas en $10^3/\mu_L$ de sangre
- NLR (X_{23}): ratio neutrófilos/linfocitos
- PLR (X_{24}): ratio plaquetas/linfocitos
- H2O2 (X_{25}): recuento de peróxido de hidrógeno en n_g/m_L de sangre
- Altura (X_{26}): altura en cm
- Peso (X_{27}): peso en Kg
- Colesterol (X_{28}): recuento de colesterol en m_g/d_L de sangre

A mayores, se han analizado gráficamente las variables *subtipo*, *grado*, *estadio*, *morfología*, *nodos* y *tamaño*, referidas únicamente a los casos de la muestra.

E.1. Librerías utilizadas y lectura de los datos

En este primer apartado de código, se presentan los paquetes de RStudio utilizados para leer y trabajar con los datos de la muestra. La base de datos donde se alojan las observaciones de todas las variables del estudio recibe el nombre de *Dataset*.

```
library(openxlsx)  # lectura de archivos xlsx
library(tidyverse) # manipulación de datos y diseño de gráficos
library(hrbrthemes) # temas de ggplot2
library(rstatix)   # función "get_summary_stats" para la manipulación y análisis de los datos
library(MASS) #cálculo de intervalos de confianza para las odds ratio
library(ROCR) # curvas ROC y AUC
library(pROC) # curvas ROC y AUC; obteción de c óptimo, sensibilidad, especificidad y precisión de los modelos
library(car)  # análisis de los factores de inflación de la varianza
library(DescTools) # cálculo de los pseudo R^2
library(glmtoolbox) # cálculo del test de Hosmer-Lemeshow
library(cvAUC) # cálculo del AUC con Validación Cruzada

Dataset <- read.xlsx("Neutrófilos_DatasetR_Nahir.xlsx", colNames = TRUE, sep.names = ",", na.strings = "NA")

head(Dataset)
dim(Dataset)
summary(Dataset)
names(Dataset)
str(Dataset)
table(Dataset$Caso)
apply(X = is.na(Dataset), MARGIN = 1, FUN = sum) # resumen de los valores perdidos
```

El siguiente código se ha utilizado para modificar la tipología de las variables, pues RStudio identificaba algunas de ellas como numéricas cuando eran categóricas, y viceversa.

```
Dataset <- Dataset %>%
mutate(Caso = as.integer(Caso)) %>%
mutate(MenoStatus = factor(MenoStatus, levels = c("0","1"), labels = c("pre", "post"),exclude = NA)) %>%
mutate(Parity = factor(Parity, levels = c("0","1", NA), labels = c("no", "si"))) %>%
mutate(BF = factor(BF, levels = c("0","1", NA), labels = c("no", "si"))) %>%
mutate(OC = factor(OC, levels = c("0","1", NA), labels = c("no", "si"))) %>%
mutate(HR = factor(HR, levels = c("0","1", NA), labels = c("no", "si"))) %>%
mutate(Smoking = factor(Smoking,levels = c("0","1", "2", NA), labels = c("nunca", "ex", "fuma"))) %>%
```

```
mutate(Stage = factor(Stage,exclude = NA)) %>%
mutate(AgeFFTP = as.numeric(AgeFFTP)) %>%
mutate(Nodes = as.integer (Nodes)) %>%
mutate(Size = as.integer(Size)) %>%
mutate(FHPS = factor(FHPS, levels = c("0","1", NA), labels = c("no", "si"))) %>%
mutate(Linfocitos_abs = as.numeric(Linfocitos_abs)) %>%
mutate(Linfocitos_perc = as.numeric(Linfocitos_perc))

summary(Dataset)
```

E.2. Justificación del modelo logístico

El código que sigue se utilizó para generar los gráficos de las Figuras 1.2 y 1.3 que justifican gráficamente porqué el modelo logístico es mejor que el lineal cuando la variable respuesta es binaria.

```
## Estudio del modelo lineal vs el modelo logístico ##

# Modelo lineal
lineal <- lm(Caso~NLR, data = Dataset)

# Figura 1.2
plot(Caso~NLR, data = Dataset, xlab = "miles por 1000/microL en sangre")
abline(lineal, col = 1, lwd=1)

# Modelo logístico

lg1 <- glm(Caso~NLR, family = binomial(link = "logit"), data = Dataset)

# Figura 1.3
plot(Caso~NLR, data = Dataset, xlab = "miles por 1000/microL en sangre")
beta1 <- coef(lg1)
curve(exp(beta1[1]+x*beta1[2])/(1+exp(beta1[1]+x*beta1[2])), add = TRUE, col = 1)
```

E.3. Curvas ROC

A continuación aparece el código utilizado para generar las Figuras 1.4 y 1.5 que sirven de apoyo a las explicaciones sobre las curvas ROC.

```
# Figura 1.4, apartados a y b.
# Ajuste del modelo de ejemplo y cálculo de predicciones
```

```
glmROC <- glm(Caso~NLR+Neutrofilos_abs+Neutrofilos_perc, family = binomial(link = "↩
    logit"), data = Dataset)
predROC <- predict(glmROC, type = "response")

Data<- Dataset[names(predROC),]
Data <- Data[, c(1,2)]
head(Data)
Data <- cbind(Data, predROC)
head(predROC)
head(Data)

# Separación entre sujetos enfermos y sanos
enfermos <- subset(Data, Caso == "1")
sanos <- subset(Data, Caso == "0")
dim(enfermos)
dim(sanos)

# Cálculo de curva ROC, sensibilidad, especificiad, AUC y c óptimo
A <- roc(Data$Caso, Data$predROC)
YoudenA <- coords(A, "best",best.method="youden", ret=c("threshold","sensitivity","↩
    specificity","accuracy")); YoudenA
A$auc

# Representación de los gráficos
# (a)
plot(density(enfermos$predROC), adjust = 1.5, xlab = "", ylab = "", main = ""
)
lines(density(sanos$predROC), adjust = 1.5)
poly_range <- density(enfermos$predROC)$x > 0 & density(enfermos$predROC)$x < ↩
    0.403266

polygon(c(0, density(enfermos$predROC)$x[poly_range], 0.403266),
c(0, density(enfermos$predROC)$y[poly_range], 0),
col = "#87CEEB",dens=20)

poly_range <- density(sanos$predROC)$x > 0.403266 & density(sanos$predROC)$x < 0.8

polygon(c(0.403266, density(sanos$predROC)$x[poly_range], 0.8),
c(0, density(sanos$predROC)$y[poly_range], 0),
col = "#87CEEB",dens=20)
segments(0.403266,0,0.403266,5,col="#B22222",lwd=2)

# (b); (a) en Figura 1.5
plot(1-A$specificities,A$sensitivities,type="S",main="", col = "#53868B",
```

```
xlab="1-especificidad",
ylab="sensibilidad", cex.lab = 1.5)
abline(a=0,b=1,lwd=1)
points(1-YoudenA$specificity,YoudenA$sensitivity,col="blue",pch=19)
text(YoudenA$specificity-0.10,YoudenA$sensitivity+0.10, "c=0.4 (0.59, 0.77)",col = "darkblue",cex=1.3)



# Figura 1.4, apartados c y d.
# Ajuste del modelo de ejemplo y cálculo de predicciones
glmSE <- glm(Caso~H2O2 + AgeDiag, family = binomial(link = "logit"), data = Dataset)
predSE <- predict(glmSE, type = "response")

Data<- Dataset[names(predSE),]
Data <- Data[, c(1,2)]
head(Data)
Data <- cbind(Data, predSE)
head(predSE)
head(Data)

# Separación entre sujetos enfermos y sanos
enfermos <- subset(Data, Caso == "1")
sanos <- subset(Data, Caso == "0")
dim(enfermos)
dim(sanos)

# Cálculo de curva ROC, sensibilidad, especificiad, AUC y c óptimo
B <- roc(Data$Caso, Data$predSE)
YoudenB <- coords(B, "best",best.method="youden", ret=c("threshold","sensitivity","specificity","accuracy"))
B$auc

# Representación de los gráficos
# (c)
plot(density(enfermos$predSE), adjust = 1.5, xlim = c(0,1), xlab = "", ylab = "", main = "")
lines(density(sanos$predSE), adjust = 1.5, xlim =c(0,1))
poly_range <- density(enfermos$predSE)$x > -1 & density(enfermos$predSE)$x < 0.46
polygon(c(0, density(enfermos$predSE)$x[poly_range], 0.46),
c(0, density(enfermos$predSE)$y[poly_range], 0),
col = "#87CEEB",dens=20)
poly_range <- density(sanos$predSE)$x > 0.46 & density(sanos$predSE)$x < 0.9
polygon(c(0.46, density(sanos$predSE)$x[poly_range], 0.9),
c(0, density(sanos$predSE)$y[poly_range], 0),
```

```
col = "#87CEEB",dens=20)
segments(0.46,0,0.46,3,col="#B22222",lwd=2)



# (d); (b) en Figura 1.5
plot(1-B$specificities,B$sensitivities,type="S",main="", col = "#53868B",
xlab="1-especificidad",
ylab="sensibilidad", cex.lab = 1.5)
abline(a=0,b=1,lwd=1)
points(1-YoudenB$specificity,YoudenB$sensitivity,col="blue",pch=19)
text(YoudenB$specificity-0.3,YoudenB$sensitivity+0.10, "c=0.46 (0.67, 0.37)",col = ↩
    "darkblue",cex=1.3)



# Figura 1.4, apartados e y f.
# Ajuste del modelo de ejemplo y cálculo de predicciones
glmROC <- glm(Caso~AgeDiag+BFMon+OC+H2O2+FHPS+NLR+Npregn, family = binomial(link = "↩
    logit"), data = Dataset)
predROC <- predict(glmROC, type = "response")

Data<- Dataset[names(predROC),]
Data <- Data[, c(1,2)]
head(Data)
Data <- cbind(Data, predROC)
head(predROC)
head(Data)

# Separación entre sujetos enfermos y sanos
enfermos <- subset(Data, Caso == "1")
sanos <- subset(Data, Caso == "0")
dim(enfermos)
dim(sanos)

# Cálculo de curva ROC, sensibilidad, especificiad, AUC y c óptimo
C <- roc(Data$Caso, Data$predROC)
YoudenC <- coords(C, "best",best.method="youden", ret=c("threshold","sensitivity","↩
    specificity","accuracy"))
C$auc

# Representación de los gráficos
# (e)
plot(density(enfermos$predROC), xlab = "", ylab = "", main = "")
lines(density(sanos$predROC))
```

```
poly_range <- density(enfermos$predROC)$x > 0 & density(enfermos$predROC)$x < 0.58

polygon(c(0, density(enfermos$predROC)$x[poly_range], 0.58),
c(0, density(enfermos$predROC)$y[poly_range], 0),
col = "#87CEEB",dens=20)

# Coloreamos FP
poly_range <- density(sanos$predROC)$x > 0.58 & density(sanos$predROC)$x < 1.2

polygon(c(0.58, density(sanos$predROC)$x[poly_range], 1.2),
c(0, density(sanos$predROC)$y[poly_range], 0),
col = "#87CEEB",dens=20)
segments(0.58,0,0.58,4,col="#B22222",lwd=2)

# (f); (c) en Figura 1.5
plot(1-C$specificities,C$sensitivities,type="S",main="", col = "#53868B",
xlab="1- especificidad",
ylab="sensibilidad", cex.lab = 1.5)
abline(a=0,b=1,lwd=1)
points(1-YoudenC$specificity,YoudenC$sensitivity,col="blue",pch=19)
text(YoudenC$specificity -0.68,YoudenC$sensitivity+0.11, "c = 0.58 (0.05, 0.81)",col =↩
    "darkblue",cex=1.3)
```

E.4. Análisis descriptivo

El siguiente código se corresponde con todos los gráficos incluidos en el apartado 2.1 del Capítulo 2, así como el estudio de la media, la desviación típica y el recuento de casos y controles de todas las variables de la base de datos analizada.

```
## Estudio exploratorio de las variables ##
casos <- subset(Dataset, Caso == "1")
controles <- subset(Dataset, Caso == "0")

summary(casos)
summary(controles)

# 1. Edad al diagnóstico (casos) o entrevista (controles).
summary(casos$AgeDiag)
summary(controles$AgeDiag)

# Hitograma y boxplot de la Figura 2.1.
par(mfrow = c(1,2))
```

```
hist(Dataset$AgeDiag, main = "", xlab = "Edad", col = "grey", freq = FALSE)
lines(density(Dataset$AgeDiag, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
boxplot(AgeDiag ~ factor(Caso,levels = c("1","0"), labels = c("Casos", "Controles")),
main = "Edad al diagnóstico o entrevista", ylab = "Edad", xlab = "", col = c("#87CEEB",
    "#FF6347") , data = Dataset)
par(mfrow = c(1,1))

# Histogramas de la Figura 2.2.
par(mfrow = c(1,2))
hist(casos$AgeDiag, main = "", xlab = "Edad en casos", col = "#FF6347", freq = FALSE)
lines(density(casos$AgeDiag, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$AgeDiag, main = "", xlab = "Edad en controles", col = "#87CEEB", freq
    = FALSE)
lines(density(controles$AgeDiag, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))

# 2. Edad a la menarquia
Dataset %>%
get_summary_stats(AgeMenarche, type = "full")

# Histograma y boxplot de la Figura 2.3.
par(mfrow=c(1,2))
hist(Dataset$AgeMenarche, main = "", xlab = "Edad a la menarquia", col = "grey", freq
     = FALSE)
lines(density(Dataset$AgeMenarche, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
boxplot(AgeMenarche ~ factor(Caso,levels = c("1","0"), labels = c("Casos", "Controles
    ")),
main = "", ylab = "Edad", xlab = "", col = c("#FF6347","#87CEEB") , data = Dataset)
par(mfrow=c(1,1))

# Histogramas de la Figura 2.4.
par(mfrow = c(1,2))
hist(casos$AgeMenarche, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$AgeMenarche, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$AgeMenarche, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$AgeMenarche, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303"
    )
par(mfrow= c(1,1))


# 3. Estado menopáusico y edad a la menopausia
tcMenoStatus <- table(Dataset$MenoStatus, factor(Dataset$Caso, levels = c("1","0"),
    labels = c("Casos", "Controles")))
Dataset %>%
```

```
get_summary_stats(AgeMenopause, type = "full")
casos %>%
get_summary_stats(AgeMenopause, type = "full")
controles %>%
get_summary_stats(AgeMenopause, type = "full")

# Diagrama de cajas y gráfico de mosaicos de la figura 2.5.
par(mfrow=c(1,2))
boxplot(AgeMenopause ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")),
main = "", xlab = "", ylab = "Edad a la menopausia", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset)
mosaicplot(tcMenoStatus, color = c("#FF6347", "#87CEEB"), main = "")
par(mfrow = c(1,1))

mosaicplot(tcMenoStatus, color = c("#FF6347", "#87CEEB"), main = "Estado menopáusico"↩
    )

# 5. Antecedentes familiares de cáncer de mama y/o ovárico
summary(Dataset$FHPS)
summary(casos$FHPS)

# Gráfico de mosaicos de a Figura 2.6
mosaicplot(tcFHPS, color = c("#FF6347", "#87CEEB"), main = "")

# 6. Número de hijos
Dataset %>%
get_summary_stats(Npregn, type = "full")
casos %>%
get_summary_stats(Npregn, type = "full")
controles %>%
get_summary_stats(Npregn, type = "full")

# Histogramas de la Figura 2.7
par(mfrow = c(1,3))
barplot(table(Dataset$Npregn), main = "", xlab = "")
barplot(table(casos$Npregn), main = "", xlab = "", col = "#FF6347")
barplot(table(controles$Npregn), main = "", xlab = "", col = "#87CEEB")
par(mfrow = c(1,1))

# 7. Edad al primer hijo nacido
Dataset %>%
get_summary_stats(AgeFFTP, type = "full")
casos %>%
```

```
get_summary_stats(AgeFFTP, type = "full")
controles %>%
get_summary_stats(AgeFFTP, type = "full")

# Histograma y gráfico de cajas de la Figura 2.8
par(mfrow= c(1,2))
hist(Dataset$AgeFFTP, main = "", xlab="", freq = FALSE)
lines(density(Dataset$AgeFFTP, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
boxplot(AgeFFTP ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles"))
    ,
xlab = "", ylab = "Edad", col = c("#FF6347", "#87CEEB"), data = Dataset)
par(mfrow= c(1,1))


# 8. Lactancia
Dataset %>%
get_summary_stats(BFMon, type = "full")
casos %>%
get_summary_stats(BFMon, type = "full")
controles %>%
get_summary_stats(BFMon, type = "full")

# Histograma y gráfico de cajas de la Figura 2.9
par(mfrow= c(1,2))
hist(Dataset$BFMon, main = "", xlab="", freq = FALSE, ylim= c(0,0.13))
lines(density(Dataset$BFMon, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
boxplot(BFMon ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "Edad", col = c("#FF6347", "#87CEEB"), data = Dataset)
par(mfrow= c(1,1))

# 10. Anticonceptivos orales y terapia hormonal substitutiva
summary(Dataset$OC)
summary(casos$OC)
summary(controles$OC)

summary(Dataset$HR)
summary(casos$HR)
summary(controles$HR)

tcOC <- table(Datasetcat$OC, factor(Datasetcat$Caso, levels = c("1","0"), labels = c(
    "Casos", "Controles")))
tcHR <- table(Datasetcat$HR, factor(Datasetcat$Caso, levels = c("1","0"), labels = c(
    "Casos", "Controles")))
mosaicplot(tcHR, color = c("#FF6347", "#87CEEB"), main = "", xlab = "Terapia hormonal
```

```
    -sustitutiva")

# Gráfico de mosaicos de la Figura 2.10
mosaicplot(tcOC, color = c("#FF6347", "#87CEEB"), main = "", xlab = "Anticonceptivos-↩
    orales")
mosaicplot(tcHR, color = c("#FF6347", "#87CEEB"), main = "", xlab = "Terapia-hormonal↩
    -sustitutiva")

# 11. Consumo de alcohol y tabaco
summary(Dataset$AlcoholFreq)
summary(casos$AlcoholFreq)
summary(controles$AlcoholFreq)

summary(Dataset$Smoking)
summary(casos$Smoking)
summary(controles$Smoking)

Alcoholcat <- cut(Dataset$AlcoholFreq, breaks = c(-Inf,0,7,Inf), labels = c("0","1-7"↩
    , ">7"),exclude = NA)
summary(Alcoholcat)

tcAlcohol <- table(Alcoholcat,factor(Datasetcat$Caso, levels = c("1","0"), labels = c↩
    ("Casos", "Controles")))
tcSmoking <- table(Datasetcat$Smoking, factor(Datasetcat$Caso, levels = c("1","0"), ↩
    labels = c("Casos", "Controles")))

# Gráficos de barras de la Figura 2.11
barplot(tcAlcohol, beside = TRUE, legend = TRUE)
barplot(tcSmoking, beside = TRUE, legend = TRUE)

# 12. Peso
Dataset %>%
get_summary_stats(Peso, type = "full")
casos %>%
get_summary_stats(Peso, type = "full")
controles %>%
get_summary_stats(Peso, type = "full")

# Histogramas y boxplot de la FIgura 2.12
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Peso, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$Peso, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Peso, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
```

```
lines(density(controles$Peso, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))

boxplot(Peso~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "Peso", col = c("#FF6347", "#87CEEB"), data = Dataset, main = "")

# 13. Altura
Dataset %>%
get_summary_stats(Altura, type = "full")
casos %>%
get_summary_stats(Altura, type = "full")
controles %>%
get_summary_stats(Altura, type = "full")

# Histogramas y boxplot de la Figura 2.14
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Altura, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$Altura, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Altura, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$Altura, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))

boxplot(Altura ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "Altura", col = c("#FF6347", "#87CEEB"), data = Dataset, main = "")

# 14. IMC
Dataset %>%
get_summary_stats(BMI, type = "full")
casos %>%
get_summary_stats(BMI, type = "full")
controles %>%
get_summary_stats(BMI, type = "full")

# Histogramas y boxplot de la Figura 2.13
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$BMI, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$BMI, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$BMI, main = "", xlab = "", col = "#87CEEB", freq = FALSE, ylim = c↩
    (0,0.086))
lines(density(controles$BMI, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))
```

```
boxplot(BMI ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "IMC", col = c("#FF6347", "#87CEEB"), data = Dataset, main = "")


# 15. Neutrófilos
Dataset %>%
get_summary_stats(Neutrofilos_abs, type = "full")
casos %>%
get_summary_stats(Neutrofilos_abs, type = "full")
controles %>%
get_summary_stats(Neutrofilos_abs, type = "full")

Dataset %>%
get_summary_stats(Neutrofilos_perc, type = "full")
casos %>%
get_summary_stats(Neutrofilos_perc, type = "full")
controles %>%
get_summary_stats(Neutrofilos_perc, type = "full")

# Histogramas y gráfico de cajas de la Figura 2.15
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Neutrofilos_abs, main = "", xlab = "", col = "#FF6347", freq = FALSE, ylim↩
    = c(0,0.35))
lines(density(casos$Neutrofilos_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303"↩
    )
hist(controles$Neutrofilos_abs, main = "", xlab = "", col = "#87CEEB", freq = FALSE, ↩
    ylim = c(0,0.4))
lines(density(controles$Neutrofilos_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "↩
    #030303")
par(mfrow = c(1,1))

boxplot(Neutrofilos_abs ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")),
xlab = "", ylab = "miles-por-1000/microL-en-sangre-", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset, main = "")


# 16. Linfocitos
Dataset %>%
get_summary_stats(Linfocitos_abs, type = "full")
casos %>%
get_summary_stats(Linfocitos_abs, type = "full")
controles %>%
```

```
get_summary_stats(Linfocitos_abs, type = "full")

# Histogramas y boxplot de la Figura 2.16
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Linfocitos_abs, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$Linfocitos_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Linfocitos_abs, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$Linfocitos_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "↩
    #030303")
par(mfrow = c(1,1))

boxplot(Linfocitos_abs ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")),
xlab = "", ylab = "miles-por-1000/microL-en-sangre-", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset, main = "")

# 17. NLR
Dataset %>%
get_summary_stats(NLR, type = "full")
casos %>%
get_summary_stats(NLR, type = "full")
controles %>%
get_summary_stats(NLR, type = "full")

# Histogramas y gráfico de cajas de la Figura 2.17
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$NLR, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$NLR, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$NLR, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$NLR, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))

boxplot(NLR ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "miles-por-1000/microL-en-sangre-", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset, main = "")



# 18. Monocitos
Dataset %>%
get_summary_stats(Monocitos_abs, type = "full")
casos %>%
get_summary_stats(Monocitos_abs, type = "full")
```

```
controles %>%
get_summary_stats(Monocitos_abs, type = "full")

# Histogramas y gráfico de cajas de la Figura 2.18
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Monocitos_abs, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$Monocitos_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Monocitos_abs, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$Monocitos_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "↩
    #030303")
par(mfrow = c(1,1))

boxplot(Monocitos_abs ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")),
xlab = "", ylab = "miles-por-1000/microL-en-sangre-", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset, main = "")

# 19. Plaquetas_abs
Dataset %>%
get_summary_stats(Plaquetas_abs, type = "full")
casos %>%
get_summary_stats(Plaquetas_abs, type = "full")
controles %>%
get_summary_stats(Plaquetas_abs, type = "full")

# Histograma y gráfico de cajas de la Figura 2.19
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Plaquetas_abs, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$Plaquetas_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Plaquetas_abs, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$Plaquetas_abs, na.rm = TRUE), lwd = 1, lty = 1, col = "↩
    #030303")
par(mfrow = c(1,1))

boxplot(Plaquetas_abs ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")),
xlab = "", ylab = "miles-por-1000/microL-en-sangre-", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset, main = "")

# 20. PLR
Dataset %>%
get_summary_stats(PLR, type = "full")
```

```
casos %>%
get_summary_stats(PLR, type = "full")
controles %>%
get_summary_stats(PLR, type = "full")

# Histogramas y gráfico de cajas de la Figura2.20
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$PLR, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$PLR, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$PLR, main = "", xlab = "", col = "#87CEEB", freq = FALSE, ylim = c(0,0.012))
lines(density(controles$PLR, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))

boxplot(PLR ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "miles-por-1000/microL-en-sangre-", col = c("#FF6347", "#87CEEB"), data = Dataset, main = "")


# 21. H2O2

Dataset %>%
get_summary_stats(H2O2, type = "full")
casos %>%
get_summary_stats(H2O2, type = "full")
controles %>%
get_summary_stats(H2O2, type = "full")

# Histogramas y gráfico de cajas de la Figura 2.21
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$H2O2, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$H2O2, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$H2O2, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$H2O2, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))

boxplot(H2O2 ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "ng/mL-", col = c("#FF6347", "#87CEEB"), data = Dataset, main = "")


# 22. Colesterol
Dataset %>%
```

```
get_summary_stats(Colesterol, type = "full")
casos %>%
get_summary_stats(Colesterol, type = "full")
controles %>%
get_summary_stats(Colesterol, type = "full")

# Histogramas y gráfico de cajas de la Figura 2.22
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Colesterol, main = "", xlab = "", col = "#FF6347", freq = FALSE, ylim = c(0,0.012))
lines(density(casos$Colesterol, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Colesterol, main = "", xlab = "", col = "#87CEEB", freq = FALSE, ylim = c(0,0.012))
lines(density(controles$Colesterol, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
par(mfrow = c(1,1))

boxplot(Colesterol ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "ng/dL-", col = c("#FF6347", "#87CEEB"), data = Dataset, main = "")

# 23. Subtipos moleculares y morfológicos de cáncer de mama de los casos de la muestra
subtype <- subset(Dataset, Subtype != "Control")
tcSubtype <- table(subtype$Subtype)
prop.table(tcSubtype)*100
labels <- c("8.27%","63.16%","13.53%","15.04%")

tcMorphology <- table(Dataset$Morphology)
round(prop.table(tcMorphology)*100,2)
labels2 <- c("80.28%", "", "-9.34%", "", "3.11%", "2.77%", "2.42%", "1.04", "")

# Gráficos de sectores de la Figura 2.23
par(mar = c(5, 0, 4, 7))
pie(tcSubtype, radius = 0.8, col = c("lightgoldenrod1", "lightblue", "mistyrose", "lightgreen"), labels = labels,
border = "white",lwd = 1)
legend(0.7,1.2, legend = c("HER2", "LuminalA", "LuminalB", "TNBC"),
fill = c("lightgoldenrod1", "lightblue", "mistyrose", "lightgreen"),
col = c("lightgoldenrod1", "lightblue", "mistyrose", "lightgreen"), bty = "n", xpd = TRUE)

par(mar = c(5, 0.3, 4, 7.5))
pie(tcMorphology, col = c("lightgoldenrod1", "#36648B", "mistyrose", "#008B45", "#
```

```
    FF8C69", "#DDA0DD", "#54FF9F",
"#C6E2FF", "#D02090"),
border = "white",lwd = 1, labels = labels2, radius = 1)
legend("topright", inset = c(-0.48, -0.2), legend = c("Ductal", "Inflammatory", "↩
    Lobular", "Medullary", "Mixed", "Mucinous",
"other", "Papillary", "Tubular"),
fill = c("lightgoldenrod1", "#36648B", "mistyrose", "#008B45", "#FF8C69", "#DDA0DD",↩
    "#54FF9F",
"#C6E2FF", "#D02090"),
col = c("lightgoldenrod1", "#36648B", "mistyrose", "#008B45", "#FF8C69", "#DDA0DD", ↩
    "#54FF9F",
"#C6E2FF", "#D02090"), bty = "n", cex = 1, xpd = TRUE)
par(mar = c(5, 4, 4, 2))

# Recuento de las variables grado, estadio, número de nodos afectados por el cáncer y↩
    tamaño. Presentes solo en los casos de la muestra

tcGrade <- table(Dataset$Grade)

tcStage <- table(Dataset$Stage)

Dataset %>%
get_summary_stats(Nodes, type = "full")

Dataset %>%
get_summary_stats(Size, type = "full")
```

E.5. Modelos de regresión ajustados

En el siguiente bloque de código, presentamos las regresiones simples de cada una de las variables de la muestra, exceptuando aquellas que sólo contienen información relativa a los casos. Los resultados de estas regresiones sentarán las bases del Modelo 1, presentado en el apartado 2.2 de Capitulo 2, y que supone el primer modelo de regresión logística considerado válido.

```
### MODELOS DE REGRESIÓN ###

## Variables por separado ##

modAgeDiag <- glm(Caso~AgeDiag,family = binomial(link = "logit"), data = Dataset, na.↩
    action=na.omit)
summary(modAgeDiag)
```

```
modAgeMenarche <- glm(Caso~AgeMenarche,family = binomial(link = "logit"), data = ↩
    Dataset, na.action=na.omit)
summary(modAgeMenarche)

modMenoStatus <- glm(Caso~MenoStatus,family = binomial(link = "logit"), data = ↩
    Dataset, na.action=na.omit)
summary(modMenoStatus)

modAgeMenopause <- glm(Caso~AgeMenopause,family = binomial(link = "logit"), data = ↩
    Dataset, na.action=na.omit)
summary(modAgeMenopause)

modParity <- glm(Caso~Parity,family = binomial(link = "logit"), data = Dataset, na.↩
    action=na.omit)
summary(modParity)
chisq.test(table(Dataset$Parity))

modNpregn <- glm(Caso~Npregn,family = binomial(link = "logit"), data = Dataset, na.↩
    action=na.omit)
summary(modNpregn)

modAgeFFTP <- glm(Caso~AgeFFTP,family = binomial(link = "logit"), data = Dataset, na.↩
    action=na.omit)
summary(modAgeFFTP)

modBF <- glm(Caso~BF,family = binomial(link = "logit"), data = Dataset, na.action=na.↩
    omit)
summary(modBF)
chisq.test(table(Dataset$BF))

modBFMon <- glm(Caso~BFMon,family = binomial(link = "logit"), data = Dataset, na.↩
    action=na.omit)
summary(modBFMon)

modBMI <- glm(Caso~BMI,family = binomial(link = "logit"), data = Dataset, na.action=↩
    na.omit)
summary(modBMI)

modOC <- glm(Caso~OC,family = binomial(link = "logit"), data = Dataset, na.action=na.↩
    omit)
summary(modOC)
chisq.test(table(Dataset$OC))

modHR <- glm(Caso~HR,family = binomial(link = "logit"), data = Dataset, na.action=na.↩
```

```
    omit)
summary(modHR)
chisq.test(table(Dataset$HR))

modAlcoholFreq <- glm(Caso~AlcoholFreq,family = binomial(link = "logit"), data = ↩
    Dataset, na.action=na.omit)
summary(modAlcoholFreq)

modSmoking <- glm(Caso~Smoking,family = binomial(link = "logit"), data = Dataset, na.↩
    action=na.omit)
summary(modSmoking)

modFHPS <- glm(Caso~FHPS,family = binomial(link = "logit"), data = Dataset, na.action↩
    =na.omit)
summary(modFHPS)
chisq.test(table(Dataset$FHPS))

modNeutrofilos_abs <- glm(Caso~Neutrofilos_abs,family = binomial(link = "logit"), ↩
    data = Dataset, na.action=na.omit)
summary(modNeutrofilos_abs)

modNeutrofilos_perc <- glm(Caso~Neutrofilos_perc,family = binomial(link = "logit"), ↩
    data = Dataset, na.action=na.omit)
summary(modNeutrofilos_perc)

modMonocitos_abs <- glm(Caso~Monocitos_abs,family = binomial(link = "logit"), data = ↩
    Dataset, na.action=na.omit)
summary(modMonocitos_abs)

modMonocitos_perc <- glm(Caso~Monocitos_perc,family = binomial(link = "logit"), data ↩
    = Dataset, na.action=na.omit)
summary(modMonocitos_perc)

modLinfocitos_abs <- glm(Caso~Linfocitos_abs,family = binomial(link = "logit"), data ↩
    = Dataset, na.action=na.omit)
summary(modLinfocitos_abs)

modLinfocitos_perc <- glm(Caso~Linfocitos_perc,family = binomial(link = "logit"), ↩
    data = Dataset, na.action=na.omit)
summary(modLinfocitos_perc)

modPlaquetas_abs <- glm(Caso~Plaquetas_abs,family = binomial(link = "logit"), data = ↩
    Dataset, na.action=na.omit)
summary(modPlaquetas_abs)
```

```
modNLR <- glm(Caso~NLR,family = binomial(link = "logit"), data = Dataset, na.action=na.omit)
summary(modNLR)

modPLR <- glm(Caso~PLR,family = binomial(link = "logit"), data = Dataset, na.action=na.omit)
summary(modPLR)

modH2O2 <- glm(Caso~H2O2,family = binomial(link = "logit"), data = Dataset, na.action=na.omit)
summary(modH2O2)

modAltura <- glm(Caso~Altura,family = binomial(link = "logit"), data = Dataset, na.action=na.omit)
summary(modAltura)

modPeso <- glm(Caso~Peso,family = binomial(link = "logit"), data = Dataset, na.action=na.omit)
summary(modPeso)

modColesterol <- glm(Caso~Colesterol,family = binomial(link = "logit"), data = Dataset, na.action=na.omit)
summary(modColesterol)
```

El siguiente apartado de código muestra los modelos de regresión presentados en el apartado 2.2 del Capítulo 2. En primer lugar, mostramos la regresión logística múltiple formada por las variables X_1, X_2, X_3, X_6, X_7, X_9, X_{11}, X_{14}, X_{15}, X_{16}, X_{17}, X_{20}, X_{21}, X_{23}, X_{24} y X_{25}, variables que resultaron significativas al ajustar los modelos de regresión simples presentados en el apartado de código anterior. Llamaremos a este Modelo 1 inicial.

```
### MODELO 1 INICIAL ##

modcomp <- glm(Caso~AgeDiag+AgeMenarche+MenoStatus+Npregn+AgeFFTP+BFMon+OC+Smoking+FHPS+Neutrofilos_abs+
Neutrofilos_perc+Linfocitos_abs+Linfocitos_perc+NLR+PLR+H2O2,family = binomial(link ="logit"), data = Dataset, na.action=na.omit)
summary(modcomp)

# Aplicamos el método del AIC
step(modcomp)
# Quitamos "Smoking", "MenoStatus", "Neutrofilos_abs", "Neutrofilos_perc" y "AgeFFTP", "FHPS", "Linfocitos_abs", "PLR", "Npregn".
```

```
modcompfinal <- glm(Caso ~ AgeDiag + AgeMenarche + BFMon + OC + NLR + Linfocitos_perc
+ H2O2, family = binomial(link = "logit"), data = Dataset,
na.action = na.omit)

summary(modcompfinal)

# Criterio AIC
step(modcompfinal)

# Análisis de los factores de inflación
vif(modcompfinal)
```

Dado que este modelo presentaba un problema de colinelidad entre las variables X_{21} y X_{23}, se valoraron los Modelos 1.A y 1.B (que posteriormente pasa a ser el Modelo 1) como alternativa. El código siguiente se corresponde con los ajustes de ambos modelos.

```
## COMPARACIÓN DE LOS MODELOS ALTERNATIVOS ###

#### MODELO 1.A ####
modfLinfo <- glm(Caso ~ AgeDiag + Linfocitos_perc
+ H2O2, family = binomial(link = "logit"), data = Dataset,
na.action = na.omit)
step(modfLinfo)
# Quitamos BFMon, AgeMenarche, OC

summary(modfLinfo)
exp(modfLinfo$coefficients)
exp(confint(object = modfLinfo, level = 0.95 ))
vif(modfLinfo)

# Pseudo R^2 y test de Hosmer-Lemeshow
PseudoR2(modfLinfo, which = "McFadden")
PseudoR2(modfLinfo, which = "McFaddenAdj")
PseudoR2(modfLinfo, which = "VeallZimmermann")
PseudoR2(modfLinfo, which = "Nagelkerke")
hltest(modfLinfo)

#### MODELO 1.B ####
modfNLR <- glm(Caso ~ AgeDiag + BFMon + OC
+ H2O2, family = binomial(link = "logit"), data = Dataset,
na.action = na.omit)
step(modfNLR)
# Quitamos NLR, AgeMenarche
```

```
summary(modfNLR)
exp(modfNLR$coef)
exp(confint(object = modfNLR, level = 0.95 ))
vif(modfNLR)


# Pseudo R^2 y test de Hosmer-Lemeshow
PseudoR2(modfNLR, which = "McFadden")
PseudoR2(modfNLR, which = "McFaddenAdj")
PseudoR2(modfNLR, which = "VeallZimmermann")
PseudoR2(modfNLR, which = "Nagelkerke")
hltest(modfNLR)
```

Como el Modelo 1.B (o Modelo 1), presenta un mejor ajuste, se escoge como alternativa al Modelo inicial, por lo que procedemos a analizar sus coeficientes $hat\beta$, las distancias de Cook, su curva ROC y su AUC.

```
# Estudio de linealidad de la Figura D.1

glm(Caso ~ AgeDiag + BFMon + OC
+ H2O2, family = binomial(link = "logit"), data = Dataset,
na.action = na.omit) %>%
augment() %>%
pivot_longer(c(AgeDiag, H2O2)) %>%
ggplot(aes(x = value, y = .fitted)) +
geom_point() +
facet_wrap(facets = vars(name), scales = "free_x") +
geom_smooth(method = "lm", formula = "y~x") +
theme_bw() +
labs(title = "", x = NULL)

# Distancias de Cook de la Figura D.2
cooks.distance(modfNLR)
windows()
plot(modfNLR)

par(mfrow = c(1,1))
plot(modfNLR, which = 4, id.n = 3)

# Predicciones
predNLR <- predict(modfNLR, type = "response")
DataNLR <- Dataset[names(predNLR),]
```

```
dim(DataNLR)
table(DataNLR$Caso)

names(predNLR) == names(predNLR)

# CURVA ROC Y AUC APARENTE
# Selección de c, precisión, sensibilidad, especificidad y Youden
LinfoNLR <- roc(DataNLR$Caso,predNLR)
res.YoudenNLR <- coords(LinfoNLR, "best",best.method="youden", ret=c("threshold","↩
    sensitivity","specificity","accuracy")); res.YoudenNLR
LinfoNLR$auc

# Cálculo del índice de Youden
youdenNLR <- (0.7142857+0.8888889)-1

# Curva ROC y AUC
ROCpredNLR <- prediction(predNLR,DataNLR$Caso)
ROCprefNLR <- performance(ROCpredNLR, "tpr", "fpr")

auc_ROCR_NLR <- performance(ROCpredNLR, measure = "auc")
auc_ROCR_NLR <- auc_ROCR_NLR@y.values[[1]]
cvAUC(predNLR,DataNLR$Caso)

# CURVA ROC y AUC CORREGIDO

# cálculo de probabilidades mediante validación cruzada
probaux=numeric(nrow(Dataset))
for (i in 1:nrow(Dataset)){
  Datasetaux=Dataset[-i,]
  mylogitaux <- glm(Caso ~ AgeDiag + BFMon + OC
  + H2O2, family = binomial(link = "logit"), data = Datasetaux,
  na.action = na.omit)
  probaux[i]=predict(mylogitaux,type=c("response"),newdata=Dataset[i,])
}
windows()
g <- roc(Dataset$Caso ~ probaux,ci=T)
plot(g,legacy.axes=T,print.auc=T, col = "#53868B")

auc(g)
ci(g) # intervalos de confianza del AUC

# cálculo del c óptimo
C <- coords(g,x="best",best.method="youden", ret=c("threshold","sensitivity","↩
    specificity","accuracy"))
```

```
# Índice de Youden
youden1 <- (0.7142857+ 0.8)-1

# Curvas ROC y AUC aparente y corregido de la Figura 2.24
#AUC aparente
plot(1-LinfoNLR$specificities ,LinfoNLR$sensitivities ,type="S",main="", col = "#53868B",
xlab="1- especificidad",
ylab="sensibilidad", cex.lab = 1)
#AUC corregido
lines(1-g$specificities ,g$sensitivities ,type="S",main="", col = "firebrick2",
xlab="1- especificidad",
ylab="sensibilidad", cex.lab = 1)
abline(a=0,b=1,lwd=1)
points(1-res.YoudenNLR$specificity ,res.YoudenNLR$sensitivity ,col="blue",pch=19)
text(res.YoudenNLR$specificity -0.750,res.YoudenNLR$sensitivity+0.10, "c = 0.604",col
= "darkblue",cex=1.3)
points(1-C$specificity ,C$sensitivity ,col="red",pch=19)
text(C$specificity -0.55,C$sensitivity -0.10, "c = 0.565",col = "darkred",cex=1.3)
legend("bottomright", legend = c("AUC aparente", "AUC corregida"), col= c("#53868B",
"firebrick2"), lty = c(1,1), bty="n")

# Representación de la matriz de confusión del AUC corregido presente en la Figura
2.25
Dataprobaux1 <- Dataset[names(probaux),]
m1 <- ifelse(probaux > 0.5646116 , "1", "0")
mean(m1 == DataNLR$Caso, na.omit = T)
table(probaux = m1, real = Dataset$Caso)



# Gráfico de cajas
boxplot(probaux~factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "Predicciones", col = c("#FF6347", "#87CEEB"), data = Dataset, main
= "")
abline(h=0.5646116,col="red")

# Gráfico de dispersión
ind <- seq(1:672)
ggplot(Dataset, aes(ind, probaux)) +
geom_point(aes(color = as.numeric(Caso)), alpha = 1, shape = 4, stroke = 1)+
geom_hline(yintercept = 0.5646116,
linetype = 2,
color = 2) +
```

```
scale_colour_gradient(low = "blue", high = "red")+
xlab("Paciente") +
ylab("Enfermedad")+
theme_bw()+
theme(legend.position = "none")
```

Añadimos también el gráfico correspondiente al estudio de la edad al diagnóstico o entrevista, variable que presentaba contradicciones con los resultados del análisis descriptivo.

```
# Estudio de la edad al diagnóstico/entrevista de la Figura A.4
casos1 <- subset(DataNLR, Caso == "1")
controles1 <- subset(DataNLR, Caso == "0")
hist(casos1$AgeDiag, xlab = "", main = "", col ="#FF6347", freq = FALSE)
lines(density(casos1$AgeDiag, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles1$AgeDiag, xlab = "", main = "", col = "#87CEEB", freq = FALSE, ylim = ↩
    c(0,0.05))
lines(density(controles1$AgeDiag, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
```

Las líneas de código que siguen, muestran los ajustes del Modelo 2, aquel que partía de variables con pocos NA's.

```
### MODELO 2: Sin variables con NA's ###
mod1 <- glm(Caso ~ AgeDiag + Parity + Npregn + BMI + Neutrofilos_perc +
Monocitos_perc
+ Colesterol, family = binomial(link = "logit"), data = Dataset,
na.action = na.omit)

summary(mod1)
vif(mod1)
step(mod1)

# Variables eliminadas: L %, AgeMenarche, L#, Mono #, NLR, Plaquetas #, N #, PLR, ↩
    MenoStatus


exp(mod1$coefficients)
exp(confint(object = mod1, level = 0.95))



# Pseudo R^2 y test de Hosmer-Lemeshow
PseudoR2(mod1, which = "McFadden")
PseudoR2(mod1, which = "McFaddenAdj")
PseudoR2(mod1, which = "VeallZimmermann")
PseudoR2(mod1, which = "Nagelkerke")
hltest(mod1)
```

```
# Estudio de la linealidad de la Figura D.3
glm(Caso ~ AgeDiag + Parity + Npregn + BMI + Neutrofilos_perc +
Monocitos_perc
+ Colesterol, family = binomial(link = "logit"), data = Dataset,
na.action = na.omit) %>%
augment() %>%
pivot_longer(c(AgeDiag, BMI, Neutrofilos_perc, Monocitos_perc)) %>%
ggplot(aes(x = value, y = .fitted)) +
geom_point() +
facet_wrap(facets = vars(name), scales = "free_x") +
geom_smooth(method = "lm", formula = "y~x") +
theme_bw()+
labs(title = "", x = NULL)

# Distancias de Cook de la Figura D.4
cooks.distance(mod1)
windows()
plot(mod1)
plot(mod1, which = 4, id.n = 3)

# Predicciones
pred1 <- predict(mod1, type = "response")

names(pred1)
Datasetpred1 <- Dataset[names(pred1),]
table(Datasetpred1$Caso)
dim(Datasetpred1)

# CURVA ROC Y AUC APARENTE
# Selección de c, precisión, sensibilidad, especificidad y Youden
B <- roc(Datasetpred1$Caso,pred1, ci=T)
res.Youden1 <- coords(B, "best",best.method="youden", ret=c("threshold","sensitivity"↩
    ,"specificity","accuracy"))

# Cálculo del índice de Youden
youden1 <- (0.6787004+0.6273885)-1

# Curva ROC y AUC
ROCpred1 <- prediction(pred1,Datasetpred1$Caso)
ROCpref1 <- performance(ROCpred1, "tpr", "fpr")

auc_ROCR1 <- performance(ROCpred1, measure = "auc")
auc_ROCR1 <- auc_ROCR1@y.values[[1]]
```

```
cvAUC(pred1,Datasetpred1$Caso)


# CURVA ROC y AUC CORREGIDO

# cálculo de probabilidades mediante validación cruzada
probaux=numeric(nrow(Dataset))
for (i in 1:nrow(Dataset)){
  Datasetaux=Dataset[-i,]
  mylogitaux <- glm(Caso ~ AgeDiag + Parity + Npregn + BMI + Neutrofilos_perc +
  Monocitos_perc
  + Colesterol, family = binomial(link = "logit"), data = Datasetaux,
  na.action = na.omit)
  probaux[i]=predict(mylogitaux,type=c("response"),newdata=Dataset[i,])
}
windows()
g <- roc(Dataset$Caso ~ probaux,ci=T)
plot(g,legacy.axes=T,print.auc=T)
auc(g)
ci(g)

# cálculo del c óptimo
C <- coords(g,x="best",best.method="youden", ret=c("threshold","sensitivity","↩
    specificity","accuracy"))

# Índice de Youden
youden2 <- (0.6570397+0.6273885)-1

# Curvas ROC y AUC aparente y corregido de la Figura 2.26
#AUC aparente
plot(1-B$specificities,B$sensitivities,type="S",main="", col = "#53868B",
xlab="1-especificidad",
ylab="sensibilidad", cex.lab = 1)
#AUC corregido
lines(1-g$specificities,g$sensitivities,type="S",main="", col = "firebrick2",
xlab="1-especificidad",
ylab="sensibilidad", cex.lab = 1)
abline(a=0,b=1,lwd=1)
points(1-res.Youden1$specificity,res.Youden1$sensitivity,col="blue",pch=19)
text(res.YoudenNLR$specificity-0.550,res.YoudenNLR$sensitivity+0.10, "c=0.456",col ↩
    = "darkblue",cex=1.3)
points(1-C$specificity,C$sensitivity,col="red",pch=19)
text(C$specificity-0.25,C$sensitivity-0.10, "c=0.462",col = "darkred",cex=1.3)
```

```
legend("bottomright", legend = c("AUC-aparente", "AUC-corregida"), col= c("#53868B", "firebrick2"), lty = c(1,1), bty="n")



# Representación de la matriz de confusión del AUC corregido presente en la Figura 2.27

# Matriz de confusión AUC corregido
Dataprobaux1 <- Dataset[names(probaux),]
m1 <- ifelse(probaux > 0.4619161 , "1", "0")
table(probaux = m1, real = Dataset$Caso)

# Gráfico de cajas de la matriz corregida
boxplot(probaux~factor(Caso, levels = c("1","0"), labels = c("Casos", "Controles")),
xlab = "", ylab = "Predicciones", col = c("#FF6347", "#87CEEB"), data = Dataset, main = "")
abline(h=0.4619161,col="red")

# Gráfico de dispersión e la matriz corregida
ind <- seq(1:672)
ggplot(Dataset, aes(ind, probaux)) +
geom_point(aes(color = as.numeric(Caso)), alpha = 1, shape = 4, stroke = 1)+
geom_hline(yintercept = 0.5646116,
linetype = 2,
color = 2) +
scale_colour_gradient(low = "blue", high = "red")+
xlab("Paciente") +
ylab("Enfermedad")+
theme_bw()+
theme(legend.position = "none")
```

Por último, incluimos los gráficos correspondientes al Apéndice A, donde se analizan las variables proporción de neutrófilos, linfocitos y monocitos en $10^3/\mu_L$ de sangre.

```


# Histogramas y gráficos de cajas de la Figura A.1
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Neutrofilos_perc, main = "", xlab = "", col = "#FF6347", freq = FALSE, ylim = c(0,0.06))
lines(density(casos$Neutrofilos_perc, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Neutrofilos_perc, main = "", xlab = "", col = "#87CEEB", freq = FALSE,
```

```
     ylim = c(0,0.06))
lines(density(controles$Neutrofilos_perc, na.rm = TRUE), lwd = 1, lty = 1, col = "↩
    #030303")
par(mfrow = c(1,1))

boxplot(Neutrofilos_perc ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")),
xlab = "", ylab = "miles por 1000/microL en sangre", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset, main = "")


# Histogramas y gráfico de cajas de la Figura A.2
Dataset %>%
get_summary_stats(Monocitos_perc, type = "full")
casos %>%
get_summary_stats(Monocitos_perc, type = "full")
controles %>%
get_summary_stats(Monocitos_perc, type = "full")

par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Monocitos_perc, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$Monocitos_perc, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303")
hist(controles$Monocitos_perc, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$Monocitos_perc, na.rm = TRUE), lwd = 1, lty = 1, col = "↩
    #030303")
par(mfrow = c(1,1))
boxplot(Monocitos_perc ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")), xlab = "", ylab = "miles por 1000/microL en sangre", col = c("#↩
    FF6347", "#87CEEB"), data = Dataset, main = "")


# Histogramas y gráfico de cajas de la Figura A.3
Dataset %>%
get_summary_stats(Linfocitos_perc, type = "full")
casos %>%
get_summary_stats(Linfocitos_perc, type = "full")
controles %>%
get_summary_stats(Linfocitos_perc, type = "full")
par(mfrow = c(1,2))
par(mar = c(5, 4, 4, 1))
hist(casos$Linfocitos_perc, main = "", xlab = "", col = "#FF6347", freq = FALSE)
lines(density(casos$Linfocitos_perc, na.rm = TRUE), lwd = 1, lty = 1, col = "#030303"↩
    )
```

```
hist(controles$Linfocitos_perc, main = "", xlab = "", col = "#87CEEB", freq = FALSE)
lines(density(controles$Linfocitos_perc, na.rm = TRUE), lwd = 1, lty = 1, col = "↩
    #030303")
par(mfrow = c(1,1))
boxplot(Linfocitos_perc ~ factor(Caso, levels = c("1","0"), labels = c("Casos", "↩
    Controles")),
xlab = "", ylab = "miles-por-1000/microL-en-sangre-", col = c("#FF6347", "#87CEEB"), ↩
    data = Dataset, main = "")
```

Bibliografía

[1] Aedo S, Pavlov S y Clavero F (2010) Riesgo relativo y Odds ratio ¿Qué son y cómo se interpretan?. Revista de Obstetricia y Ginecología, 5: 51-54. Doi:10.13140/2.1.4654.6886

[2] Aguilar JP, Arriaga MB, Chaves NM y Zeballos DR (2017) Entendiendo la Odds Ratio. Revista SCientifica, 15:27-30. http://200.7.173.107/index.php/Scientifica/article/view/38

[3] Ali AMG, Schmidt MK, Bolla MK, Wang Q, Gago-Dominguez M et al. (2014) Alcohol Consumption and Survival after a Breast Cancer Diagnosis: A Literature-Based Meta-analysis and Collaborative Analysis of Data for 19.239 Cases. Cancer Epidemiology, Biomarkers & Prevention, 23:934-945. https://doi.org/10.1158/1055-9965.EPI-13-0901

[4] American Cancer Society (2019) Comprensión de un diagnóstico de cáncer de seno. Estado del receptor hormonal del cáncer de seno. https://www.cancer.org/es/cancer/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-seno/estado-del-receptor-hormonal-del-cancer-de-seno.html. Accedido el 8 de febrero de 2023.

[5] Bonita R, Beaglehole R y Kjellström T (2008) Epidemiología básica. Organización Panamericana de la Salud, 629: 1-279. https://iris.paho.org/handle/10665.2/3153

[6] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA et al. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. A Cancer Journal for Clinicians, 68: 394-424. https://doi.org/10.3322/caac.21492

[7] BREOGAN (2018) BREast Oncology Galician Network. https://proyectobreogan.es/

[8] van Buuren S y Groothuis-Oudshoorm K (2011) mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45 (3): 1-67 https://www.jstatsoft.org/article/view/v045i03

[9] Cárdenas J (2015) Odd Ratio: qué es y cómo se interpreta. Networkianos. Blog de Sociología. https://networkianos.com/odd-ratio-que-es-como-se-interpreta/. Accedido el 15 de febrero de 2023.

[10] CDC (2022) Alcohol and Public Health. Center of Disease Control and Prevention. https://www.cdc.gov/alcohol/faqs.htm#23heavyDrinking

[11] Cecchini RS, Costantino JP, Cauley JA, Cronin WN, Wickerham DL et al. (2012) Body mass index and the risk for developing invasive breast cancer among high-risk women in NSABP P-1 and STAR breast cancer prevention trials. Cancer Prevention Research, 5: 583-59. https://doi.org/10.1158/1940-6207.CAPR-11-0482

[12] Cerda J, Vera C y Rada G (2013) Odds ratio: aspectos teóricos y prácticos. Revista médica de Chile, 141: 1329-1335. http://dx.doi.org/10.4067/S0034-98872013001000014

[13] Chen W, Rosner B, Hankinson SE, Colditz GA y Willett WC (2011) Moderate Alcohol Consumption Durin Adult Life, Drinking Patterns, and Breast Cancer Risk. JAMA, 306: 1884-1890. Doi: 10.1001/jama.2011.1590

[14] Cruz GI, Martínez ME, Natarajan L, Wertheim BC, Gago-Dominguez M et al. (2013) Hypothesized role of pregnancy hormones on HER2+ breast tumor development. Breast Cancer Research and Treatment, 137: 237-246. https://doi.org/10.1007/s10549-012-2313-0

[15] DePolo J (2022). Subtipos moleculares de cáncer de mama. BREASTCANCER.ORG. https://www.breastcancer.org/es/tipos/subtipos-moleculares. Accedido el 16 de abril de 2023.

[16] DePolo J (2023). Carcinoma ductal invasivo (CDI). BREASTCANCER.ORG. https://www.breastcancer.org/es/tipos/carcinoma-ductal-invasivo. Accedido el 16 de abril de 2023.

[17] DePolo J (2023). Carcinoma lobular invasivo (CLI). BREASTCANCER.ORG. https://www.breastcancer.org/es/tipos/carcinoma-lobular-invasivo. Accedido el 17 de abril de 2023.

[18] DePolo J (2023). Cáncer de mama inflamatorio (CMI). BREASTCANCER.ORG.https://www.breastcancer.org/es/tipos/inflamatorio. Accedido el 17 de abril de 2023.

[19] Dolle JM, Daling JR, White E, Brinton LA, Doody DR et al. (2009) Risk factors for triple-negative breast cancer in women under the age of 45 years. Cancer Epidemiology, Biomarkers & Prevention, 18: 1157-1166. https://doi.org/10.1158/1055-9965.EPI-08-1005

[20] Ethier JL, Desautels D, Templeton A, Shah PS y Amir E (2017) Prognostic role of neutrophil-to-lymphocyte ratio in breast cancer: a systematic review and meta-analysis. Breast Cancer Research, 19: 2. https://doi.org/10.1186/s13058-016-0794-1

[21] Fan J, Upadhye S y Worster A (2015) Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine, 8: 19-20. https://doi.org/10.1017/S1481803500013336

[22] Fanjul A (2020) Non Parametric Methods for the Comparison of ROC Curves with Application to Biomedicine. Tesis, Universidad de Santiago de Compostela. https://minerva.usc.es/xmlui/handle/10347/25152?show=full

[23] Faria SS, Fernandes PC, Barbosa MJ, Lima VC, Fontes W et al. (2016) The neutrophil-to-lymphocyte ratio: a narrative review. Ecancermedicalscience, 10:702. https://doi.org/10.3332/ecancer.2016.702

[24] Gago-Domingez M, Castelao JE, Gude F, Peña M, Guado-Barrera ME et al. (2016) Alcohol and breast cancer tumor subtypes in a Spanish Cohort. SpringerPlus, 5:39. https://doi.org/10.1186/s40064-015-1630-2

[25] Gago-Domíngez M, Matabuena M, Redondo CM, Patel SP, Carracedo A et al. (2020) Neutrophil to lymphocyte ratio and breast cancer risk: analysis by subtype and potential interactions. Scientific Reports, 10: 13203. https://doi.org/10.1038/s41598-020-70077-z

[26] Gago-Domingez M, Redondo CM, Calaza M, Matabuena M, Bermudez MA et al. (2021) LIPG endothelial lipase and breast cancer risk by subtypes. Scientific Reports, 11:10436. https://doi.org/10.1038/s41598-021-89669-4

[27] Hosmer DW, Lemeshow S y Sturdivant RX (2013) Applied Logistic Regression. John Wiley & Sons, New Jersey. Recuperado de https://doi-org.ezbusc.usc.gal/10.1002/9781118548387.ch4

[28] Instituto Nacional del Cáncer. https://www.cancer.gov/espanol. Accedido el 8 de febrero de 2023.

[29] Ipaguirre A, Barrio I y Rodríguez- Álvarez MX (2019) On the optimism correction on the area under the receiver operating characteristic curve in logistic prediction models. SORT, 43 (1): 145-162. https://www.idescat.cat/sort/sort431/43.1.7.iparragirre-etal.pdf

[30] Iwase T, Sangai T, Sakakibara M, Sakakibara J, Ishigami E et al. (2016) An increased neutrophil-to-lymphocyte ratio predicts poorer survival following recurrence for patients with breast cancer. Molecular and Clinical Oncology, 6: 266-270. https://doi.org/10.3892/mco.2016.1101

[31] Jiang X, Castelao JE, Chavez-Uribe E, Fernández B, Celeiro C et al. (2012) Family History and Breast Cancer Hormone Receptor Status in a Spanish Cohort. PLoS ONE, 7: e29459. https://doi.org/10.1371/journal.pone.0029459

[32] Kabat GC, Heo M, Kamensky V, Miller AB y Rohan TE (2012) Adult Height in relation to risk of cancer in a cohort of Canadian women. International Journal of Cancer, 132: 1125-1132. https://doi.org/10.1002/ijc.27704

[33] Kawai M, Malone KE, Tang MC y Li CI (2014) Height, body mass index (BMI), BMI change, and the risk of estrogen receptor-positive, HER2-positive, and triple-negative breast cancer among women ages 20 to 44 years. Cancer, 120: 1548-1556. https://doi.org/10.1002/cncr.28601

[34] Lawrence MH (2006) Logistic Regression: An Overview. Eastern Michigan University, College of Technology. https://www.researchgate.net/publication/255597415_Logistic_Regression_An_Overview

[35] Ma H, Wang Y, Sullivan-Halley J, Weiss L, Marchbanks PA et al. (2010) Use of Four Biomarkers to Evaluate the Risk of Breast Cancer Subtypes in the Women's Contraceptive and Reproductive Experiences Study. Cancer Research, 70: 575-587. https://doi.org/10.1158/0008-5472.CAN-09-3460

[36] Mirón JA y Alonso M (2008) Medidas de Frecuencia, Asociación e Impacto en Investigación Aplicada. Medicina y seguridad del Trabajo, 211: 93-102. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0465-546X2008000200011

[37] Palmer JR, Boggs DA, Wise LA, Ambrosone CB, Adams-Campbell LL et al. (2011) Parity and Lactation in Relation to Estrogen Receptor Negative Breast Cancer in African American Women. Cancer Epidemiology, Biomarkers & Prevention, 20: 1883-1891. Doi:https://doi.org/10.1158/1055-9965.epi-11-0465

[38] Pharoah PD, Day NE, Duffy S, Easton DF y Ponder BAJ (1997) Family history and the risk of breast cancer: A systematic review and meta-analysis. International Journal of Cancer, 71: 800-809. https://doi.org/10.1002/(SICI)1097-0215(19970529)71:5<800::AID-IJC18>3.0.CO;2-B

[39] Phipps AI, Chlebowski RT, Prentice R, McTiernan A, Wactawski-Wende J et al. (2011) Reproductive History and Oral Contraceptive Use in Relation to Risk of Triple-Negative Breast Cancer. Journal of the National Cancer Institute, 103: 470-477. https://doi.org/10.1093/jnci/djr030

[40] Picon-Ruiz M, Morata-Tarifa C, Valle-Goffin JJ, Friedman ER y Slingerland JM (2017) Obesity and adverse breast cancer risk and outcome: Mechanistic insights and strategies for intervention. CA: A Cancer Journal for Clinicians, 67: 378-397. https://doi.org/10.3322/caac.21405

[41] Pita S y Pértegas S (2003) Pruebas diagnósticas: Sensibilidad y especificidad. Cadernos de Atención Primaria, 10(1), 120-124. https://www.fisterra.com/formacion/metodologia-investigacion/pruebas-diagnosticas-sensibilidad-especificidad/

[42] R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[43] Real Academia Española (2023). https://www.rae.es/. Accedido el 6 de febrero de 2023.

[44] Reeves MJ, Newcomb PA, Remington PL, Marcus PM y MacKenzie WR (1996) Body mass and breast cancer: Relationship between method of detection and stage of disease. American Cancer Society, 77: 301-307. https://doi.org/10.1002/(SICI)1097-0142(19960115)77:2<301::AID-CNCR12>3.0.CO;2-5

[45] Redondo CM, Gago-Domínguez M, Miranda S, Enguix M, Jiang X et al. (2012) Breast Feeding, Parity and Breast Cancer Subtypes in a Spanish Cohort. PLoS ONE, 7: e40543. https://doi.org/10.1371/journal.pone.0040543

[46] Ritte R, Lukanova A, Tjonneland A, Olsen A, Overvad K et al. (2012) Heigh, age at menarche and risk of hormone receptor-positive and-negative breast cancer: A cohort study. International Journal of Cancer, 132: 2619-2629. https://doi.org/10.1002/ijc.27913

[47] Rubin DB (1987) Introduction. En Multiple Imputation for Nonresponse in Surveys: 1-23. John Wiley & Sons, Canada.

[48] Sheater SJ (2009) A modern approach to regression with R. Springer.

[49] Smith TJ y McKenna CM (2013) A Comparision of Logistic Regression Pseudo R^2 Indices. Multiple Linear Regression Viewpoints, 39(2): 17-26. https://www.semanticscholar.org/paper/A-Comparison-of-Logistic-Regression-Pseudo-R-2-Smith-McKenna/6e21d24f52645f9e871f6896a6bf13267d2ad7f0

[50] Stoltzfus JC (2011) Logistic Regression: A brief primer. Academic Emergency Medicine, 18:1099-1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

[51] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I et al. (2021) Blogal Cancer Statistics 2020: GLOBOCAN Estimates of Incedence and Mortality Worldwide for 36 Cancer in 185 Countries. American Cancer Society, 71: 209-249. https://doi.org/10.3322/caac.21660

[52] Suzuki R, Orsini N, Saji S, Key TJ y Wolk A (2009) Body weight and incidence of breast cancer defined by estrogen and progesterone receptor status. A meta-analysis. International Journal of Cancer, 124: 698-721. https://doi.org/10.1002/ijc.23943

[53] Szumilas M (2010) Explaining Odds Ratios. Journal of the Canadian Academy of Child and Psychiatry, 19: 227-229. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

[54] Trentham-Dietz A, Newcomb PA, Storer BE, Longnecker MP, Baron J et al. (1997) Body Size and Risk of Breast Cancer. American Journal of Epidemiology, 145: 1011-1019. https://doi.org/10.1093/oxfordjournals.aje.a009057

[55] Uribe-Querol E y Rosales C (2015) Neutrophils in Cancer: Two Sides of the Same Coin. Journal of Inmunology Research, 2015: 1-21. https://doi.org/10.1155/2015/983698

[56] Yang S y Berdine G (2017) The reciever operating characteristic (ROC) curve. The Southwest Respiratory and Critical Care Chronicles, 5(19): 34-36 https://www.researchgate.net/publication/316751328_The_receiver_operating_characteristic_ROC_curve

[57] Zhou X, Du Y, Huang Z, Xu J, Qiu T et al. (2014) Prognostic value of PLR in Varius Cancer: A Meta-Analysis. PLoS ONE, 9: https://doi.org/10.1371/journal.pone.0101119

[58] Zhang B, Shu X, Delahanty RJ, Zheng C, Michailidou K et al. (2015) Height and Breast Cancer Risk: Evidence From Prospective Studies and Mendelian Randomization. Journal of the National Cancer Institute, 107: djv219. https://doi.org/10.1093/jnci/djv219