TRABAJO FIN DE MASTER EN BIOESTADÍSTICA

REGRESIÓN LOGÍSTICA EN AUSENCIA DE GOLD STANDARD DESDE LA PERSPECTIVA BAYESIANA

SEPTIEMBRE 2021

Virginia Rafael Núñez

Tutoras: Julia Amador Pacheco y Rosario Susi García

ÍNDICE

Listado de tablas:

Listado de figuras:

RESUMEN

Cuando se analiza una variable respuesta dicotómica (Sí/No) desde la perspectiva frecuentista siendo esta una prueba diagnóstica pero no es Gold Standard se presenta el problema de cómo introducir la incertidumbre de la sensibilidad y especificidad de dicha prueba. Para este problema se presenta un modelo de regresión logística desde la perspectiva bayesiana con el fin de estimar no solo los coeficientes del modelo sino, también, la especificidad y sensibilidad de dicha prueba ya que son desconocidas.

Por otro lado, como frecuentemente es complicado introducir la información a priori acerca de los coeficientes del modelo, en el modelo realizado se utiliza un método alternativo de tal modo que la información a priori que se incorpora de estos es a través de las medias condicionales a priori (CMP de sus siglas en inglés). Estas son las probabilidades de obtener una respuesta exitosa en diversas combinaciones de las covariables del modelo.

Se analizó un modelo principal (modelo 1) el cual consideraba sensibilidad y especificidad desconocidas y se incorporaba la información a priori de estas a través de distribuciones beta y la información a priori de las covariables a través de distribuciones beta de las CMP. Adicionalmente se analizó el modelo con diferentes distribuciones a priori de los parámetros y desde la perspectiva frecuentista con el fin de comparar los resultados. Todos los modelos se realizaron sobre una base de datos simulada.

Los resultados obtenidos mostraron que los dos modelos con distribuciones a priori informativas para sus coeficientes del modelo (modelo 1 con sensibilidad y especificidad desconocidas y modelo 3 con sensibilidad y especificidad iguales a 1) tenían un mejor ajuste que aquellos con distribuciones a priori no informativas (modelo 2 con sensibilidad y especificidad desconocidas y modelo 4 con sensibilidad y especificidad iguales a 1): $DIC_{modelo\ 1}$=321.2 y $DIC_{modelo\ 3}$=319.8 frente a $DIC_{modelo\ 4}$=321.7 y la no convergencia de las cadenas del modelo 2. En relación a las estimaciones de los coeficientes y de la mayoría de los ORs los modelos 1 y 3 presentaron los intervalos de probabilidad más ajustados frente a los intervalos de probabilidad del modelo 4 y los intervalos de confianza del modelo frecuentista. Para el modelo 1, además, todos los coeficientes y ORs obtenidos fueron relevantes.

Palabras clave: regresión logística, sensibilidad, especificidad, estadística bayesiana, medias condicionales a priori.

ABSTRACT

When we analyze a dichotomous response variable (Yes/No) from a frequentist perspective, being a diagnostic test but not Gold Standard, it poses the problem of how to introduce the uncertainty of the sensitivity and specificity of the test. To address this issue, we present a logistic regression model from a Bayesian perspective, in order to estimate not only the coefficients of the model but also the specificity and sensitivity of the test, which are unknown.

At the same time, as it is often complicated to introduce a priori information about the model coefficients, an alternative method is also used in this model, so that a priori information is incorporated through the conditional means priors (CMP). These are the probabilities of obtaining a successful response in various combinations of the model covariates.

We have analyzed a main model (model 1), which considered unknown sensitivity and specificity and incorporated the a priori information through beta distributions and the prior information of the covariates through beta distributions of the CMPs. Additionally, the model was analyzed with different prior distributions of the parameters and from a frequentist perspective, in order to compare the results obtained. All models were run on a simulated database.

The results obtained showed that the two models with informative prior distributions for their model coefficients (model 1 with unknown sensitivity and specificity, and model 3 with sensitivity and specificity equal to 1) had a better fit than those with non-informative prior distributions (model 2 with unknown sensitivity and specificity, and model 4 with sensitivity and specificity equal to 1): $DIC_{model\ 1}$=321.2 y $DIC_{model\ 3}$=319.8, against $DIC_{model\ 4}$=321.7 and the non-convergence of the chains of model 2. Regarding the coefficient estimates and most of the ORs, models 1 and 3 presented the best-fit probability intervals, against the probability intervals of model 4 and the confidence intervals of the frequentist model. In addition, for model 1, all the coefficients and ORs obtained were relevant.

Keywords: logistic regression, sensitivity, specificity, Bayesian statistics, conditional means priors.

1 INTRODUCCIÓN

En ciencias de la salud es muy común estudiar la relación que existe entre una variable respuesta de tipo binario y unas variables explicativas (independientes) cuantitativas y/o cualitativas. Por ejemplo, relación entre dejar de fumar y factores demográficos y basales como la edad, el nivel de estudio y el número de cigarrillos iniciales en mujeres embarazadas (1), o relación entre alcanzar un objetivo de HbA1c (hemoglobina glicosilada) y factores clínicos y no clínicos en pacientes con diabetes Mellitus tipo II (2). Este tipo de relación se estudia mediante la regresión logística binaria.

Este trabajo evaluará el primer ejemplo expuesto anteriormente, la relación entre dejar de fumar (Sí/No) y las covariables edad, el nivel de estudios y el número de cigarrillos iniciales en mujeres embarazadas. Lo datos serán simulados y posteriormente analizados y se basan en un ensayo clínico realizado a mujeres embarazadas fumadoras en Baltimore, Maryland, en el año 1984 (3). El ensayo se diseñó para evaluar como una reducción en el consumo de tabaco durante el embarazo podía aumentar el peso de los neonatos. El ensayo era aleatorizado a dos brazos, tratamiento y control. El tratamiento consistía en recibir una intervención para dejar de fumar por parte de dos profesionales, una con experiencia en el asesoramiento sobre el embarazo y la otra con experiencia sobre el tabaquismo. La intervención se realizaba, principalmente, por teléfono, correo y alguna presencialmente, y consistía en estimular a las mujeres para dejar de fumar proporcionándoles información, apoyo, orientación práctica y estrategias de comportamiento. Las mujeres aleatorizadas al grupo control no recibieron ninguna intervención para dejar de fumar. Entre otras, las características basales que se recogieron fueron la edad, los años de estudio y el número de cigarrillos diario antes de la aleatorización. En el octavo mes de embarazo se preguntó a las embarazadas su nivel de tabaquismo, es decir, el número de cigarrillos diarios que fumaban.

En relación a la variable respuesta se considera, No, si a los 8 meses de embarazo las mujeres fumaban algún cigarrillo diario y, Sí, si no fumaban ningún cigarrillo diario. Analizando la naturaleza de esta variable, se puede decir que es una prueba para clasificar si una mujer fuma o no y, como consecuencia, es una forma de medir si el programa de intervención ha funcionado o no, pero lo cierto es que tiene un carácter subjetivo. La asunción estándar en el análisis de regresión binaria es considerar que la variable respuesta es medida de forma perfecta, es decir, se podría suponer que todas las mujeres son sinceras y que cuando se les preguntó por el número de

cigarrillos diarios que fumaban al octavo mes de embarazo no mintieron. La realidad es que no hay forma de saberlo, dado que las mujeres estaban llevando su vida cotidiana y nadie estaba con ellas durante la realización del estudio.

Esto presenta una las claves de este trabajo, la ausencia de Gold Standard, porque no existe una prueba que garantice a un 100% de probabilidad si una mujer ha dejado de fumar o no, o dicho de otra forma, no es muy preciso preguntar a las mujeres si han dejado de fumar o no. Este problema no es único y se presenta en muchas otras pruebas diagnósticas de clasificación (4). Por ejemplo, las pruebas diagnósticas basadas en imágenes, pruebas diagnósticas para la detección de la cardiopatía isquémica, pruebas rápidas para la faringitis estreptocócica o los test de laboratorio (5), (6), (7), (8).

Traduciéndolo a términos estadísticos, esto se corresponde con la sensibilidad y especificidad de una prueba o test diagnóstico. Estos términos hacen referencia a los verdaderos positivos y verdaderos negativos, respectivamente, es decir, la capacidad de una prueba de clasificar como enfermos o que existe cierta condición a los sujetos realmente enfermos o que realmente la tienen y, la capacidad de clasificar como sanos o que no existe cierta condición a los sujetos realmente sanos o que realmente no la tienen (9). Por lo tanto, la sensibilidad y especificidad de la variable "dejar fumar" son desconocidas.

Un camino por el que se puede abordar este problema de una forma sencilla es desde la perspectiva bayesiana ya que permite asumir que ambos parámetros son desconocidos e incluir información a priori acerca de estos parámetros. Por ejemplo, incluyendo información de profesionales de cuánto de probable es que una mujer embaraza que ha dejado de fumar realmente diga que ha dejado de fumar, y cuánto de probable es que una mujer embaraza que no ha dejado de fumar realmente diga que no ha dejado de fumar. De igual modo, se puede incluir información a priori acerca de los coeficientes de las covariables que se quieren estimar.

Por lo tanto, en este trabajo se realizará un modelo de regresión logística binaria cuando la variable respuesta es resultado de una prueba diagnóstica que no es Gold Standard con sensibilidad y especificidad desconocidas.

Para ello, en primer lugar será necesario simular la base de datos dado que no se dispone de la base de datos original. Esta estará formada por las covariables edad, nivel de estudios y estatus de

tabaquismo previo, y la variable respuesta dejar de fumar (Sí/No). Para simular las covariables se recopilará toda la información descriptiva disponible en el artículo de referencia (1) y los dos artículos que le preceden (3), (10) y, para la variable respuesta se recopilarán las estimaciones de los coeficientes de las covariables estimados en el artículo (1). Una vez simulada la base de datos se realizará un análisis descriptivo y se compararán los resultados con los datos reales de los artículos de referencia. El siguiente paso será encontrar las distribuciones a priori para los parámetros del modelo, que son la sensibilidad, la especificidad y los coeficientes de regresión; para ello habrá que elicitar los parámetros de dichas distribuciones a priori. Todo esto se recoge en la sección 3 de este TFM. A continuación, en la sección 4, se analizará el modelo desde la perspectiva bayesiana y se analizarán otros modelos desde la perspectiva bayesiana cambiando alguna asunción y por último, se analizará un modelo desde la perspectiva frecuentista. Por último, se comparan los resultados obtenidos en los diferentes modelos.

2 OBJETIVOS

El objetivo principal es ajustar un modelo de regresión logística desde la perspectiva bayesiana teniendo en cuenta la ausencia de un Gold Standard e incorporando información a priori informativa de los parámetros del modelo (sensibilidad, especificidad a través de distribuciones a priori betas y los coeficientes de las covariables mediante las CMP). Este se considerará el modelo principal.

Como objetivos secundarios, para poder llevar a cabo el principal, son:

- Simulación de una base de datos de acuerdo a la información que se presenta en el artículo de referencia de McIntur et al. de 2004 (1) y los artículos en los cuales se basa este (3), (10).
- Determinación de las distribuciones a priori de los coeficientes mediante las CMP y elicitación de los hiperparámetros de las distribuciones a priori betas de la sensibilidad, especificidad y de las CMP.

Otros objetivos secundarios serían:

- Ajuste de un modelo de regresión logística desde la perspectiva bayesiana asumiendo sensibilidad y especificidad de igual modo que el modelo principal y distribuciones a priori normales no informativas para los parámetros del modelo.
- Ajuste de un modelo de regresión logística desde la perspectiva bayesiana asumiendo que la variable respuesta es el Gold Standard (es decir, sensibilidad y especificidad iguales a 1) y las CMP para los parámetros del modelo.
- Ajuste de un modelo de regresión logística desde la perspectiva bayesiana asumiendo que la variable respuesta es el Gold Standard y distribuciones a priori normales no informativas para los parámetros del modelo.
- Ajuste de un modelo de regresión logística asumiendo que la variable respuesta es el Gold Standard desde la perspectiva frecuentista.
- Comparación de los coeficientes y OR estimados en los cinco modelos

3 METODOLOGÍA

En esta sección se resumirá la metodología usada y el porqué de ajustar un modelo de regresión logística desde la perspectiva bayesiana en ausencia de Gold Standard. También se explicará el procedimiento para obtener una base de datos simulada y la forma de introducir información a priori de los coeficientes a través de las CMP.

3.1 SIMULACIÓN BASE DE DATOS

La necesidad de simular una base de datos surge cuando no se dispone de una base de datos real que se ajuste al problema que se va estudiar.

Como se ha mencionado anteriormente, este trabajo se basa en el artículo de McInturff et al. de 2004 (1), que a su vez, se basa en otros dos artículos el de Sexton et al. de 1987 (3) y el de Magder et al. de 1997 (10), donde se pueden encontrar características de la base de datos real. Se usará toda la información disponible en ellos acerca de las covariables (edad, nivel de estudios y número previo de cigarrillos diario) y de la variable respuesta (dejar de fumar) para construir una base de datos lo más similar posible a la base de datos original.

En este trabajo la simulación se basará en:

1. Identificar las partes que cambian aleatoriamente y describirlas como variables aleatorias.

2. Generar las variables aleatorias.

3. Validar el resultado comparándolo con la información real disponible.

1. Identificar las partes que cambian aleatoriamente y describirlas como variables aleatorias

Para simular una base de datos similar a la base de datos real en la cual se basa una investigación es necesario elaborar un modelo que se ajuste con la situación real.

Para ello, es necesario conocer que partes cambian aleatoriamente. Estas partes que cambian son las variables que se recogen en un estudio o ensayo clínico y toman valores diferentes dependiendo de cada sujeto investigado.

Existen dos tipos de variables a simular:

- Variables explicativas
- Variables dependientes

Para cualquiera de los dos tipos de variables es necesario conocer si se dispone de información útil que pueda usarse para reconocer su naturaleza.

Esta información útil se puede encontrar en los resúmenes descriptivos, resultados y conclusiones de artículos de referencia. Con esta información se puede conocer si una variable es de carácter cuantitativo o cualitativo (o categórico).

Para las variables de tipo cuantitativo será útil conocer la media y la desviación típica (o varianza), en cambio, para las variables de tipo cualitativo será la frecuencia absoluta y el porcentaje sobre el total que se ha calculado. Otra información interesante serían las tablas de contingencia entre dos o más variables de tipo categórico, las cuales, aportan información acerca de la asociación entre ellas.

Con todo esto, se puede intuir si las variables siguen alguna distribución conocida que luego será usada para simularlas.

2. Generar las variables aleatorias

Conocida la naturaleza de las variables a simular y, suponiendo que estas siguen alguna distribución conocida, el siguiente paso es generarlas.

Existen diferentes distribuciones conocidas que se pueden usar fácilmente para simular, como son: la normal, binomial, uniforme, exponencial, beta, geométrica, etc.

Distribuciones de las variables explicativas (covariables)

Para simular las covariables, se centrará la atención en dos distribuciones conocidas que serán las que se apliquen en este trabajo:

- Distribución Normal (μ, σ) donde el parámetro, μ, es la media y, σ, la desviación típica de la variable a simular.
- Distribución de Bernoulli (p), donde el parámetro, p, es la probabilidad de éxito de una característica.

Como se indica, para utilizar estas distribuciones es necesario conocer ciertos parámetros de las mismas. Estos serán obtenidos de la información útil que se extraiga de los artículos de referencia.

La distribución Normal se usará para la simulación de las covariables cuantitativo y la Bernoulli para el cálculo de las covariables categóricas dicotómicas.

Con el programa R (11), esta tarea será llevada a cabo con las siguientes funciones desarrolladas para tal fin:

- rnorm(n, mean, sd) donde, n, corresponde con el número de observaciones que se quiere simular y, mean y sd, corresponden con la media y desviación típica de la variable.
- rbern(n, p) donde, n, corresponde con el número de observaciones que se quiere simular y, p, corresponde a la probabilidad de éxito de la variable.

Al generar de esta forma las covariables, se asume, indirectamente, que estas son independientes, es decir, que no existen asociaciones entre ellas. Por otro lado, también se obtendrán valores de las covariables que, quizás, estén fuera del rango de los valores observados en la investigación.

Para hacer frente a ello, es necesario generar inicialmente una muestra lo suficientemente grande para después depurar la base de datos eliminando las observaciones fuera de rango y eliminando los registros que tengan una combinación de covariables imposibles.

Una vez depurada la base de datos, como último paso, se obtendrá una muestra de esta con el número de sujetos que se desee analizar.

Cálculo de la variable respuesta

Una forma de obtener la variable respuesta dicotómica (dejar de fumar: Sí/No) es con las estimaciones de los coeficientes del modelo de regresión logística obtenidos en el artículo McInturff et al. de 2004 (1).

Estos coeficientes serán utilizados para calcular las probabilidades estimadas asociadas a la variable respuesta de la siguiente forma:

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2 + \cdots + \hat{\beta}_m * X_k)}{1 + (\exp(\hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2 + \cdots + \hat{\beta}_m * X_k))} \quad [1]$$

donde:

$X_1, X_2, \dots, X_k$: corresponde a las covariables

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m$: corresponden a los coeficientes estimados asociados a las covariables

Como la variable respuesta no es Gold Standard, esta probabilidad se puede reescribir en función de la sensibilidad y especificidad del modo siguiente:

$$\hat{q} = \hat{p}\, Se + (1 - \hat{p})(1 - Sp) \quad [2]$$

Para poder calcular esta probabilidad será necesario dar unos valores a la sensibilidad y especificidad. Estos valores serán las modas de las distribuciones a priori de cada uno de ellos y serán obtenidos igualmente del artículo de referencia.

Una vez obtenidas estas probabilidades para cada sujeto se simulará la variable respuesta aplicando la distribución bernoulli para cada registro asumiendo un ensayo con probabilidad igual a la probabilidad estimada obtenida anteriormente.

3. Comparar los resultados simulados con la información real disponible

Como último paso, una vez obtenida la base de datos simulada, se realizará un análisis descriptivo de esta y se compararán los resultados con los obtenidos de los artículos. De este modo, se puede chequear que ambas tienen las mismas características.

3.2 REGRESIÓN LOGÍSTICA

La regresión logística es un tipo de análisis donde se relaciona una variable respuesta Y de tipo categórico con unas variables independientes cuantitativas y/o cualitativas. Esta variable dependiente que se quiere predecir puede ser medida en una escala binaria, nominal u ordinal. Por ejemplo, presencia o ausencia de cierta enfermedad (o dicho de otro modo, éxito o fracaso de cierto resultado), gravedad de una enfermedad, etc. Este trabajo se centra en la regresión logística donde la variable respuesta es binaria (Dejar de fumar: Sí/No).

Se denota la variable aleatoria binaria como:

$$Y = \begin{cases} 1, & \text{si el resultado es un éxito,} \\ 0, & \text{si el resultado es un fracaso,} \end{cases}$$

cuyas probabilidades son: $\Pr(Y = 1) = \pi$ y $\Pr(Y = 0) = 1 - \pi$

Si se tienen n variables aleatorias tales como esta $Y_1, \dots, Y_n$ la cuales son independientes con $\Pr(Y_j = 1) = \pi_j$, entonces la probabilidad conjunta es (12):

$$\Pr(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{j=1}^{n} \pi_j^{y_j} \left(1 - \pi_j\right)^{1-y_j} = exp\left[\sum_{j=1}^{n} y_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log\left(1 - \pi_j\right)\right]$$

siendo $\{y_1, \dots, y_n\} \in \{0,1\}$

Por otra parte lo que interesa es estudiar la relación entre una o más variables independientes o explicativas, $\boldsymbol{X}$ (así se denotará el vector de covariables por simplicidad) y la variable Y. El modelo logístico establece la siguiente relación entre la probabilidad de que el resultado sea un éxito, dado que el individuo presenta los valores $X_1 = x_1, X_2 = x_2, \dots, X_k = x_k$:

$$\Pr(Y = 1|\boldsymbol{X}) = \frac{1}{1+exp^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} = \frac{1}{1+exp^{-(\boldsymbol{X}^T\beta)}} = \frac{exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}}{1+exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}} \qquad [3]$$

Por lo tanto, la probabilidad de que sea un fracaso será:

$$\Pr(Y=0|\boldsymbol{X}) = 1 - \Pr(Y=1|\boldsymbol{X}) = 1 - \frac{1}{1+exp^{-(\boldsymbol{X}^T\boldsymbol{\beta})}} = \frac{exp^{-(\boldsymbol{X}^T\boldsymbol{\beta})}}{1+exp^{-(\boldsymbol{X}^T\boldsymbol{\beta})}} = \frac{1/exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}}{1+1/exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}}$$

$$= \frac{1}{1+exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}} \quad [4]$$

Dado que el modelo de regresión logística se enmarca dentro del conjunto de Modelos Lineales Generalizados (GLM) (12) y la variable dependiente sigue una distribución de bernoulli, la cual pertenece a la familia de distribuciones exponenciales, la probabilidad $\Pr\big((Y=1|\boldsymbol{X})\big) = \pi$ se pueden expresar como:

$$g(\pi) = \boldsymbol{X}^T\boldsymbol{\beta}$$

donde:

$\boldsymbol{X}$: es el vector de variables explicativas

$\boldsymbol{\beta}$: es el vector de parámetros a estimar

g : es la función de enlace

Existen diferentes funciones de enlace (probit, log log, logit) pero en este caso, como se trata de regresión logística, la función que aplica es la *logit*, de tal forma que:

$$logit(\pi) = log\Big(\frac{\pi}{1-\pi}\Big) = logit\big(Pr(Y=1|\boldsymbol{X})\big) = log\left(\frac{Pr(Y=1|\boldsymbol{X})}{1-Pr(Y=1|\boldsymbol{X})}\right)$$

Aplicando [3] y [4] se obtiene que:

$$log\left(\frac{Pr(Y=1|\boldsymbol{X})}{1-Pr(Y=1|\boldsymbol{X})}\right) = \log\left(\frac{\frac{exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}}{1+exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}}}{\frac{1}{1+exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}}}\right) = \log\big(exp^{(\boldsymbol{X}^T\boldsymbol{\beta})}\big) = \boldsymbol{X}^T\boldsymbol{\beta}$$

Esta otra forma de expresar la relación [3] cuando se usa la regresión logística es el denominado odds, es decir, el cociente entre la probabilidad de que el resultado sea un éxito y la probabilidad de que sea un fracaso. Con los odds se puede calcular el denominado Odds Ratio (OR) de una covariable fijados unos valores del resto de las covariables $\boldsymbol{X}$ y el cual se denota como $OR(\boldsymbol{X})$. Este se define como la razón de odds, es decir, el cociente entre el odds en el grupo de presenta una categoría de la covariables y el odds en el grupo que no presenta esa categoría. . El OR, dado que es un cociente con $OR \in \{0, \infty\}$, se utiliza para interpretar el parámetro en el que se basa en términos de asociación con la ocurrencia del evento, fijados el resto de parámetros. Cuando el OR>1 la asociación es positiva, es decir, la presencia de cierta categoría de la covariable se asocia con una mayor ocurrencia del evento. En caso de que OR<1 entonces la asociación es negativa y en caso que OR=1 entonces se interpreta como que no hay asociación.

Dependiendo de si la covariable es cuantitativa o cualitativa la interpretación del OR es ligeramente diferente. De forma que el OR del parámetro β asociado a una covariable X cuantitativa se interpreta como la razón de tener un éxito respecto de que no es tantas veces menor o mayor (OR<1 o OR>1 respectivamente) por cada unidad que aumenta la covariable X. Mientras que cuando la covariable X es cualitativa el OR se interpreta como la razón de tener un éxito respecto de que no es tantas veces menor o mayor (OR<1 o OR>1 respectivamente) de la categoría 1 de la covariable X en comparación con la categoría 2 de la misma covariable.

Una vez introducido el modelo de regresión logística, el siguiente paso es estimar los parámetros del modelo ($\boldsymbol{\beta}$). En lo explicado anteriormente solo se ha hecho referencia a los parámetros de los coeficientes del modelo, es decir, de las covariables. Sin embargo, como se comentó en la introducción de este trabajo la variable respuesta que se analizará tiene la característica de ser resultado de una prueba diagnóstica con sensibilidad y especificidad desconocidas. Por lo tanto, estos dos parámetros habrá que introducirlos en el modelo y posteriormente también estimarlos.

3.3 SENSIBILIDAD Y ESPECIFICADAD

Se considera una prueba dicotómica usada para diagnosticar cierta enfermedad o presentar cierta condición D. Se denota como Z como el estado verdadero de la enfermedad o condición entonces se denota:

$Z = 0$, cuando D no está presente,

$Z = 1$, cuando D está presente.

Por otro lado, se denota como Y el resultado de la prueba diagnóstica donde:

$Y = 0$, cuando la prueba indica que D no está presente.

$Y = 1$, cuando la prueba indica que D está presente.

Se definen sensibilidad y especificidad de la prueba diagnóstica como:

Sensibilidad: probabilidad de una prueba diagnóstica de clasificar que existe cierta condición en un individuo que realmente la tiene.

$$Se = \Pr(Y = 1|Z = 1)$$

Especificidad: probabilidad de una prueba diagnóstica de clasificar que no existe cierta condición en un individuo que realmente no la tiene.

$$Sp = \Pr(Y = 0|Z = 0)$$

En este trabajo se asume que la sensibilidad y la especificidad de la prueba no dependen de ninguna covariable en el modelo y se asume que la clasificación errónea del estado de la condición no es diferencial con respecto a las covariables.

Usando la regresión binomial para modelizar la probabilidad de que un individuo presente cierta condición D en función de unas covariables $\boldsymbol{X}$ quedaría:

$$\pi = \Pr(Z = 1|\boldsymbol{X})$$

Como se ha explicado antes, en el caso de la regresión logística π sería:

$$\pi = \frac{\exp(\boldsymbol{X}^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^T\boldsymbol{\beta})}$$

Aplicando la ley de la probabilidad total se tiene que la probabilidad de obtener un resultado positivo en la prueba, dadas unas covariables es:

$$q_x = \Pr(Y=1|\boldsymbol{X}) = \pi\, Se + (1-\pi)(1-Sp) \qquad [5]$$

Una vez introducidos los parámetros de sensibilidad y especificidad al modelo de probabilidad el siguiente paso es estimarlos. Esto se abordará desde la perspectiva bayesiana y se explicará en las siguientes secciones.

3.4 INFERENCIA DESDE LA PERSPECTIVA FRECUENTISTA

Como se ha comentado en la sección 3.2 una vez planteado el modelo de regresión que se va a usar el siguiente paso es estimar los parámetros del modelo usando los datos observados. El método más común es el método de máxima verosimilitud (MLE por sus siglas en inglés).

MLE consiste en encontrar los valores de los parámetros que maximizan la función de verosimilitud, siendo ésta la función de masa conjunta de la muestra vista como función de los parámetros. Esta función se denota como $\mathcal{L}(parámetro/s)$ donde normalmente al parámetro/s a estimar se les denota como θ.

Se denota como $\mathcal{L}(\beta, (Y, \boldsymbol{X}))$ a la función de verosimilitud asociada a una muestra de n individuos, para un modelo de regresión logística con parámetros $\boldsymbol{\beta}$ asociados a las variables explicativas $\boldsymbol{X}$ y con una variable respuesta dicotómica Y. Entonces se tiene que la función de verosimilitud es:

$$\mathcal{L}(\boldsymbol{\beta}, (Y, \boldsymbol{X})) = \prod_{j=1}^{n} P\left(Y_j = y_j \middle| \boldsymbol{X}_j\right)^{y_j} \left(1 - P\left(Y_j = y_j \middle| \boldsymbol{X}_j\right)\right)^{1-y_j}$$

donde:

$y_j \in \{0,1\}$, $\boldsymbol{X}_j = \left(X_{1j}, X_{2j}, \dots, X_{kj}\right)$, para $j = 1, \dots, n$.

A continuación se expresan las probabilidades $P(Y_j = y_j | \boldsymbol{X}_j)$ en términos de especificidad y sensibilidad aplicando [3]:

$$P(Y_j = 1 | \boldsymbol{X}_j) = q_{xj}$$

$$P(Y_j = 0 | \boldsymbol{X}_j) = 1 - q_{xj}$$

Por lo tanto, la función de verosimilitud en términos de sensibilidad, especificidad quedaría como:

$$\mathcal{L}(\boldsymbol{\beta}, Se, Sp, (Y, \boldsymbol{X})) = \prod_{j=1}^{n} \left(q_{xj}\right)^{y_j} (1 - q_{xj})^{1-y_j} =$$

$$= \prod_{j=1}^{n} \left(\pi_j \, Se + \left(1-\pi_j\right)(1-Sp)\right)^{y_j} \left(1 - \left(\pi_j \, Se + \left(1-\pi_j\right)(1-Sp)\right)\right)^{1-y_j}$$

$$= \prod_{j=1}^{n} \left(\frac{\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)}{1-\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)} \, Se + \left(1 - \frac{\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)}{1-\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)}\right)(1-Sp) \right)^{y_j} \left(1 \right.$$

$$\left. - \left(\frac{\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)}{1-\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)} \, Se + \left(1 - \frac{\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)}{1-\exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_j)}\right)(1-Sp) \right) \right)^{1-y_j}$$

Como se puede aprecia los parámetros a estimar son la sensibilidad, especificidad y los coeficientes de las covariables (dado que π depende de los β como se indica en [2] y [3]). Según el artículo de Magder et al. de 1997 (10) se puede estimar con el MLE la sensibilidad y especificidad de forma simultánea a los coeficientes del modelo pero presenta un gran problema: el modelo está saturado de parámetros a estimar $(Se, Sp, \beta_0, \beta_1, \ldots, \beta_k)$ y hay un número infinito de estimaciones de máxima verosimilitud, es decir, hay infinitas soluciones. Una solución que se da en ese artículo es asumir ciertos valores para la sensibilidad y especificidad. El problema de esta solución es que en muchas ocasiones los valores de la sensibilidad y especificidad son desconocidos y deben ser estimados, tal y como sucede en el caso de este TFM.

Por lo tanto, el enfoque frecuentista no da solución al problema que se plantea en este trabajo y habrá que usar otra metodología para inferir los parámetros del modelo.

3.5 INFERENCIA DESDE LA PERSPECTIVA BAYESIANA

3.5.1 Concepto

El enfoque bayesiano en el análisis de datos es, cada vez, más usado, concretamente en ciencias de la salud, dado que aporta unos resultados libres de los inconvenientes del tamaño muestral y las pruebas de significación, tan presentes en el análisis estadístico frecuentista (13). Por otro lado, permite meter información a priori de los parámetros a estimar.

Mientras que, desde la perspectiva frecuentista, se estudia la probabilidad de los datos supuestas las hipótesis del estudio como ciertas, en la estadística bayesiana se estudia la probabilidad de las hipótesis que se planteen en un estudio, es decir, se mide la certeza o no sobre estas dados los datos y la información a priori.

Por ejemplo, supongamos que se realiza un estudio clínico para evaluar si la edad es un factor importante para describir la ocurrencia de desarrollar cierta patología (o evento) o no, dado un tratamiento. La hipótesis nula sería que la edad no afecta al desarrollo de la patología, frente a la hipótesis alternativa que sería que si afecta. Desde el punto de vista frecuentista, la pregunta sería, cuánto de probable es tener unos datos tan extremos o más que se recogen en el estudio asumiendo que la edad no es un factor importante para desarrollar la patología. En cambio, desde la perspectiva bayesiana, la pregunta sería, cuánto de probable es que la edad no sea un factor importante teniendo los datos recogidos del estudio.

En el libro de Gelman et al. de 2014 (14) se define como la característica principal de los métodos bayesianos al uso explícito de la probabilidad para cuantificar la incertidumbre en las inferencias basadas en los datos. Dicho de otro modo, es posible calcular las probabilidades, no solo, de la variable respuesta en estudio, si no, también, las probabilidades de los factores de interés, hipótesis y modelos.

En este libro se establece que el proceso del análisis de datos bayesiano se basa en 3 pasos:

- Establecer un modelo de probabilidad total: una distribución de probabilidad conjunta para todas las cantidades observables (datos) y no observables en un problema.
- Condicionamiento a los datos observados: cálculo e interpretación de la distribución a posteriori y comprobación de si esta es apropiada.

- Evaluación del ajuste del modelo y las implicaciones de la distribución posterior resultante.

De forma general, en la metodología bayesiana se denota como $Datos$ a un conjunto de datos, y θ al vector de parámetros desconocidos a estimar.

Antes de hacer formulaciones de la probabilidad de θ dados unos datos es necesario empezar con un modelo que proporcione la distribución de probabilidad conjunta de θ y los datos. Para ello es necesario recurrir al teorema de Bayes:

De forma general, sea $\{A_1, A_2, \dots, A_n\}$ un conjunto de sucesos excluyentes con probabilidades distintas de 0 y sea B otro suceso con probabilidad distinta de 0 tal que se conocen las probabilidades condicionales $P(B|A_i)$ entonces la probabilidad condicional de $P(A_i|B)$ es:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \qquad [6]$$

De forma sencilla se reemplaza A_i por el vector de parámetros θ y B por $Datos$, entonces según la definición de probabilidad condicionada:

$$P(\theta, Datos) = P(\theta)P(Datos|\theta)$$

donde $P(\theta)$ es la distribución a priori y $P(Datos|\theta)$ es la distribución de los datos.

Aplicando Bayes [6] se tiene que

$$P(\theta|Datos) = \frac{P(Datos|\theta)P(\theta)}{P(Datos)} \qquad [7]$$

donde $P(\theta|Datos)$ es la distribución a posteriori y $P(Datos)$ es la distribución marginal de los datos, es decir, la distribución predictiva a priori de los datos.

De forma equivalente a [7] omitiendo el factor $P(Datos)$ dado que no depende de θ, el conjunto de datos es fijo y, por lo tanto, se puede considerar como constante se tiene que:

$$P(\theta|Datos) \propto P(Datos|\theta)P(\theta) \qquad [8]$$

La función de $P(Datos|\theta)$ es llamada función de verosimilitud $\mathcal{L}(Datos|\theta)$, por lo tanto, la expresión [8] se puede escribir como:

$$P(\theta|Datos) \propto \mathcal{L}(Datos|\theta)P(\theta)$$

Estas fórmulas comprenden la clave de la inferencia Bayesiana, la distribución a posteriori es proporcional al producto de la función de verosimilitud y la distribución a priori. Esta última es la que incorpora la información previa que se tenga de los parámetros a estimar. Para introducir esta información previa es necesario recurrir a distribuciones conocidas. Estas dependen a su vez de parámetros o hiperparámetros. Al proceso de dar valores a estos hiperparámetros se le llama elicitación y es lo que se explica en la sección siguiente.

Una vez obtenida la distribución a posteriori, esta será la que proporcionará toda la información para hacer las inferencias sobre los parámetros. Para cada parámetro ($Se, Sp, \boldsymbol{\beta}$) se obtendrán las estimaciones puntuales de la media, desviación típica, error estándar y los percentiles 2.5, 25, 50, 75, 97,5.

Por otro lado, se obtendrá con los percentiles los llamados intervalos de probabilidad o credibilidad, los cuales se interpretan como, por ejemplo, la probabilidad de que el valor de θ esté dentro del intervalo es del 0.95 (o con una confianza del 95% el intervalo contiene al parámetro).

3.5.2 Distribuciones a priori – Elicitación

Existen diversas formas de introducir la información a priori que se tenga de los parámetros que posteriormente se estimarán. En este trabajo se van a usar dos métodos. Para la sensibilidad y especificidad se utilizará la distribución $Beta(\alpha,\beta)$, cuyos parámetros se elicitarán, es decir, se le darán valores de acuerdo a la información a priori de que se disponga.

Por otro lado, para los coeficientes asociados a las covariables se usará un método para obtener las medias condicionales a priori para coeficientes de regresión binomial.

3.5.2.1 Distribución Beta

Para estimar la Se y Sp que, como se han descrito anteriormente son dos probabilidades, incorporando información previa, la función $Beta$ es una buena opción ya que es una distribución lo suficientemente flexible para modelizar las creencias que se tengan sobre estos dos parámetros.

La función $Beta(\alpha,\beta)$ depende de dos escalares de forma, ambos son mayores que 0 ($\alpha,\beta>0$).

Sea θ una variable aleatoria (Se o Sp en este caso), α,β dos escalares ($\alpha,\beta>0$) y θ sigue una distribución $Beta$ ($\theta \sim Beta(\alpha,\beta)$) entonces se tiene que la función de densidad $Beta(\alpha,\beta)$ de θ es:

$$p(\theta)=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\quad \theta\in[0,1]\,,$$

donde $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, y cuya media, varianza y moda son:

$Media=\frac{\alpha}{\alpha+\beta}$, $Varianza=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, $Moda=\frac{\alpha-1}{\alpha+\beta-2}$.

Esta función toma valores entre [0-1] al igual que la Se y Sp, por lo tanto, es una buena elección de distribución a priori de dichos parámetros.

El siguiente paso es la elicitación de los parámetros de la distribución $Beta(\alpha, \beta)$, es decir, transformar la opinión de los expertos en una distribución $Beta$ concreta. Existen diversas maneras de buscar estos valores:

- A partir de dos momentos
- Con el valor de dos cuantiles
- Con la moda y un cuantil
- Con la moda y la media
- A partir de los datos de otro caso similar

Por ejemplo, si se tiene creencia del valor de la media y la moda de la sensibilidad, estos se sustituirían en las formulas anteriores y despejando se obtendría los valores de α y β.

3.5.2.2 Medias condicionales a priori (CMP)

Una forma sencilla de introducir información previa de los coeficientes asociados a las covariables (β) sería a través de la distribución $Normal$ dado que estos pueden tomar valores reales. Sin embargo, en esta sección se aplica un método alternativo donde la información a priori que se introduce no es directamente de los parámetros sino de las medias de las combinaciones de estos cuando son binomiales. En el artículo de Bedrick et al. de 1996 (15) se describe que dado que la media de una variable dicotómica es la probabilidad de éxito entonces, para cada localización, se puede especificar la distribución de la media de las posibles observaciones en dicha localización, refiriéndose como localización a cada combinación de valores de las covariables. Este método se denomina medias condicionales a priori. A partir de ahora este método se denotará como CMP por simplificar la notación.

Dado que introducir información previa de los coeficientes puede resultar difícil este método introduce la información de los expertos especificando las probabilidades de que cierta condición D esté presente (en este caso dejar de fumar) dadas diversas combinaciones de las covariables. Dicho de otro modo, las a priori para los coeficientes $\boldsymbol{\beta}$ se obtendrán a partir de las distribuciones a priori de estas probabilidades o medias condicionales.

Antes de empezar, especificar, que para usar este método las covariables que se analicen tienen que estar recogidas de forma categórica. En este trabajo serán de la siguiente forma:

- Edad recogida en 3 categorías: <20, [20-30], >30 años
- Nivel de estudios recogida en 3 categorías: Menos de educación secundaria, educación secundaria y estudios universitarios.
- Estatus de tabaquismo previo recogida en 2 categorías: <20, ≥ 20 cigarrillos diarios

Se consideran k coeficientes de regresión $(\beta_0, \beta_1, \dots, \beta_k)$ y se seleccionan las k combinaciones de covariables $(\tilde{x_i}, i = 1, \dots, k)$ linealmente independientes dentro del rango de las posibles. Entonces se introduce la información previa de las probabilidades de éxito dadas estas combinaciones en vez directamente los coeficientes β, de tal forma que $\tilde{\pi_i}, i = 0,1, \dots, k$ son las probabilidades a priori de éxito dadas estas combinaciones.

Para introducir el conocimiento previo, dado que se debe introducir información de las probabilidades de éxito bajo un conjunto de combinaciones de covariables se usarán diferentes distribuciones $Beta(\alpha, \beta)$, una para cada una de las combinaciones que se han seleccionado, de igual modo que se ha explicado en la sección anterior.

Por lo tanto, con las seleccionadas k combinaciones independientes de covariables se obtiene una transformación biyectiva entre el vector de coeficientes $\boldsymbol{\beta}$ y las $\tilde{\pi} = (\widetilde{\pi_0}, \widetilde{\pi_1}, \dots, \widetilde{\pi_k})'$ tal que:

$$\boldsymbol{\beta} = \widetilde{\boldsymbol{X}^{-1}} g(\tilde{\pi})$$

donde: $\widetilde{\boldsymbol{X}} = (\widetilde{x_0}', \dots, \widetilde{x_k}')'$

3.5.3 Modelo estadístico

El objetivo de este estudio es modelizar la relación entre la variable dicotómica dejar de fumar (Sí/No) y las covariables categóricas: edad, nivel de estudios y estatus de tabaquismo previo teniendo la sensibilidad y especificidad de la variable respuesta desconocidas y queriendo incorporar información previa de expertos en el modelo en relación a estos parámetros y coeficientes de las covariables.

Para modelizar esta relación se utilizará un modelo de regresión binomial, concretamente el modelo de regresión logística.

Se denota como $(Y_j, \boldsymbol{X}_j), j = 1, \dots, n$ los datos observados de una muestra de tamaño n, siendo Y_j la variable respuesta binaria (resultado de una prueba diagnóstica) y $\boldsymbol{X}_j$ el vector fila de las covariables del individuo j. Por otro lado, se denota como Z_j al estado real de la variable respuesta. Se asume que $Y_j|\boldsymbol{X}_j \sim Bernoulli(q_j)$ donde la probabilidad de éxito está definida por $q_j = Pr(Y_j = 1 \mid \boldsymbol{X_j}) = \pi_j Se + (1 - \pi_j)(1 - Sp)$ como se explicó en la sección 3.3. Se tiene también que la probabilidad real de éxito de un individuo j dada su información de las covariables $\boldsymbol{X}_j$ es $\pi_j = \Pr(Z_j = 1|\boldsymbol{X}_j)$. Por lo tanto, la función de verosimilitud de los parámetros queda como el producto de probabilidades de la forma:

$$\mathcal{L}(Se, Sp, \beta) = \prod_{j=1}^{n} [\pi_j Se + (1 - \pi_j)(1 - Sp)]^{y_j} \, [\pi_j (1 - Se) + (1 - \pi_j) Sp]^{1 - y_j}$$

Como distribuciones a priori de Se, Sp y los CMPs se asumen $Betas$ independientes, por lo tanto, la distribución a posteriori quedaría como:

$$p(Se, Sp, \beta | X, Y) \propto \mathcal{L}(Se, Sp, \beta) p(Se) p(Sp) p(\beta)$$

donde: $p(\beta)$ se puede expresar en función de las a priori de las probabilidades de éxito de las distintas combinaciones de las covariables:

$$p(\beta) = c \prod_{i=0}^{k} \tilde{\pi}_i^{a_i - 1} (1 - \tilde{\pi}_i)^{b_i - 1} |J|_+$$

donde $|J|_+ = \prod_{i=0}^{k} \tilde{\pi}_i (1 - \tilde{\pi}_i) \left|\widetilde{\boldsymbol{X}}\right|_+$, $c = \prod_{i=0}^{k} \Gamma(a_i + b_i) / \{\Gamma(a_i)\Gamma(b_i)\}$ y (a_i, b_i) corresponden a los hiperparámetros de las distribuciones beta $[Be(a_i, b_i)]$ asociadas a cada una de las combinaciones de covariables.

Debido a la complejidad del modelo es imposible llegar a la expresión matemática exacta para la distribución a posteriori, por lo que será necesario recurrir a métodos de simulación

computacionales para generar muestras de distribuciones de probabilidad. El método que se usará se explicará en la sección siguiente.

3.5.4 Métodos de Cadenas de Markov Monte Carlo (MCMC)

MCMC viene del inglés Monte Carlo Markov Chain que significa Cadenas de Markov de Monte Carlo. Los MCMC son métodos de simulación para generar muestras de distribuciones de probabilidad basados en la construcción de cadenas de Markov. Estos métodos son muy usados en la metodología bayesiana cuando la forma exacta de la distribución a posteriori de los parámetros es desconocida obteniendo muestras de la distribución a posteriori mediante la construcción de una cadena de Markov cuya distribución estacionaria sea $p(\theta|\boldsymbol{X})$.

Resumidamente las cadenas de Markov en tiempo discreto es un proceso estocástico donde la probabilidad de un evento depende únicamente del evento anterior. Se construye una cadena de Markov $\theta^{(1)}$, $\theta^{(2)}$,... cuya distribución estacionaria sea $p(\theta|\boldsymbol{X})$. Cada vector simulado $\theta^{(t)}$ depende del vector previo $\theta^{(t-1)}$ y se simula a partir de la distribución de transición $q(.|\theta^{(t-1)})$. Finalmente la cadena convergerá a $p(\theta|\boldsymbol{X})$ independiente de sus valores iniciales.

Para este proceso es necesario simular varias cadenas con un número alto de iteraciones para aproximarse a la distribución estacionaria. De estos, los primeros vectores serán descartados (este proceso se denomina burnin) y el resto formaran la muestra de la distribución estacionaria con la cual se aproximarán las inferencias a posteriori de interés.

Este método computacional se llevará a cabo mediante la librería BRugs (16) del programa R (11) el cual es flexible para el análisis bayesiano de modelos estadísticos complejos utilizando algoritmos MCMC.

Cuando se utilizan estos métodos es necesario comprobar la convergencia de las cadenas tras el burnin, esto se analizará de forma gráfica mediante las gráficas de trazas. En estas se muestran por colores los valores simulados de las cadenas en cada iteración y si no se observan tendencias y las trazas se superponen entonces las cadenas convergen y si cada una tiene una dirección (es decir, presentan desviaciones de estacionaridad) entonces no convergen. En el eje x se mostrarán las

iteraciones a lo largo del tiempo y en el eje y se mostrarán los valores simulados de las cadenas para un parámetro. Se obtendrá una gráfica de trazas por cada uno de los parámetros a estimar.

Otra forma de comprobar la convergencia de las cadenas será de forma analítica mediante los criterios de convergencia de Geweke y PSRF (Potencial Scale Reduction Factor). El criterio de Geweke se basa en la comparación de medias de dos partes de la cadena (por ejemplo, un porcentaje del principio de la cadena y porcentaje del final de la cadena). Si las muestras se extraen de una distribución estacionaria entonces las dos medias son iguales y el estadístico de Geweke tiene una distribución asintóticamente normal estándar, es decir, el valor del estadístico Z pertenece al intervalo (-1.96; 1.96).

Por otro lado, el criterio de convergencia de PSRF se basa en la comparación de la varianza dentro de las cadenas con la varianza entre cadenas. Si las cadenas convergen el valor de PSRF estará próximo a 1, y cualquier desviación de la igualdad sugerirá que las cadenas aún tienen que converger.

Por último, con este método también se obtiene el DIC (Deviance Information Criteria) el cual es una medida para comparar modelos, muy similar al AIC.

El DIC se basa en el deviance que se define como:

$$deviance(\theta) = -2\log(\mathcal{L}(\theta, \boldsymbol{X}))$$

Puesto que el deviance depende de la verosimilitud, por lo tanto, a mayor verosimilitud de los datos menor deviance y, por lo tanto, menor DIC y mejor ajuste del modelo.

4 RESULTADOS

Los resultados que se muestran en este epígrafe se dividen en dos partes, la primera corresponde a las consideraciones y obtención de la base de datos simulada y, la segunda muestra los resultados (coeficientes y ORs estimados) obtenidos en los diferentes modelos planteados en los objetivos.

4.1 SIMULACIÓN DE LA BASE DE DATOS

4.1.1 Extracción de la información de los artículos de referencia

La base de datos simulada consta de la variable respuesta dejar de fumar y tres covariables: edad, nivel de estudios y estatus de tabaquismo previo. Todas las variables eran de tipo cualitativo de la forma:

Variable respuesta: Dejar de fumar: Sí/No

Covariables:

- Edad: <20, [20-30], >30 años
- Nivel de estudios: Menos de educación secundaria, educación secundaria y estudios universitarios
- Estatus de tabaquismo previo: <20, ≥ 20 cigarrillos diarios

En primer lugar, fue necesario extraer información útil de los artículos de referencia acerca del tamaño de la muestra y estas variables.

Como se explicó en la sección 3.1, para las covariables cuantitativas fue necesario encontrar en los resúmenes descriptivos el valor de la media y desviación típica y para las cualitativas fue necesario encontrar las frecuencias absolutas y relativas. Aunque todas las covariables eran de tipo cualitativo, es decir, tienen categorías, existía la posibilidad de que la información que se proporcionara en los artículos acerca de ellas fuera de tipo de cuantitativo en vez de cualitativo. Esto no fue un problema dado que se usó la información de la media y desviación típica para simular la covariable y luego esta fue categorizada.

Para la variable respuesta, se utilizaron los coeficientes estimados del modelo, es decir, los $\hat{\beta}$, siendo:

$\widehat{\beta_0}$: el coeficiente asociado al término independiente

$\widehat{\beta_1}$: el coeficiente asociado a la categoría de [20-30] años de edad

$\widehat{\beta_2}$: el coeficiente asociado a la categoría de >30 años de edad

$\widehat{\beta_3}$: el coeficiente asociado a la categoría de educación secundaria en nivel de estudios

$\widehat{\beta_4}$: el coeficiente asociado a la categoría de estudios universitarios en nivel de estudios

$\widehat{\beta_5}$: el coeficiente asociado a la categoría de < 20 cigarrillos diarios en estatus de tabaquismo previo

A continuación, se muestra en la Tabla 1 la información recogida en los artículos que posteriormente se usó para simular la base de datos.

Tabla 1 Información descriptiva para la simulación de la base de datos

Artículo de referencia	Parámetro	Información útil
McInturff et al. de 2004 (1)	Tamaño de la muestra	$n = 361$
	Coeficientes estimados de las covariables del modelo bayesiano con sensibilidad y especificidad desconocidos	$\hat{\beta}_0 = -1.27$ $\hat{\beta}_1 = -1.75$ $\hat{\beta}_2 = -1.32$ $\hat{\beta}_3 = 0.86$ $\hat{\beta}_4 = 0.37$ $\hat{\beta}_5 = 1.29$
	Moda de la sensibilidad y especificidad	$Moda_{Se} = 1$ $Moda_{Sp} = 0.93$

Sexton et al. de 1987 (3)	Media y desviación típica de la edad en años y el nivel de estudios recogido como número de años estudiados. Rango de años de la edad	$Edad \in [14-42]$ $años$: $\mu = 24.9;\ \sigma = 4.7$ $Nivel\ de\ estudios$: $\mu = 12.3;\ \sigma = 2.0$
Magder et al. de 1997 (10)	Número y porcentaje de pacientes que fumaban menos de un paquete diario y número y porcentaje de ellas que fumaba más	$Estatus\ previo\ de\ tabaquisco$: $< 20\ cigarrillos/día$: $n(\%) = 254\ (70.36\%)$ $\geq 20\ cigarrillos/día$: $n(\%) = 107\ (29.64\%)$

Como se puede ver en la tabla se disponía de información sobre el tamaño de la muestra, los coeficientes estimados y la frecuencia absoluta y relativa de la covariable status de tabaquismo previo. Sin embargo, la información disponible de las covariables edad y nivel de estudios fue de tipo cuantitativo, es decir, solo se disponía de la media y la desviación típica. Por otro lado, tampoco se encontró información cruzada entre las covariables, por ejemplo, media y desviación típica de la edad o el nivel de estudios según el estatus de tabaquismo previo, etc.

4.1.2 Simulación de las covariables de la base de datos

Para la simulación fue necesario empezar con un tamaño de muestra lo suficientemente grande como para poder depurar la base de datos posteriormente de las combinaciones de covariables imposibles. Por lo tanto, el tamaño de muestra inicial que se asumió fue de 10000 mujeres.

Las covariables se simularon de la siguiente forma:

- Edad: se simularon 10000 registros bajo una normal de media 24.9 y desviación típica 4.7
- Nivel de estudios: se simularon 10000 registros bajo una normal de media 12.3 y desviación típica 2.0

- Estatus de tabaquismo previo: se simularon 10000 registros bajo una Bernoulli con probabilidad de 0.2964 de pertenecer a la categoría de mujeres que fumaban ≥ 20 cigarrillos diarios.

A todas variables se les aplicó un redondeo para no tener números decimales.

4.1.3 Depuración de la base de datos y categorización de las covariables

El siguiente paso fue limpiar la base de datos con solo las covariables eliminando registros donde alguna covariable estuviera fuera de rango o alguna combinación de covariables careciera de sentido.

Dado que de los artículos se pudo extraer el rango de edad de las mujeres estudiadas, inicialmente se eliminaron todos los registros donde la edad de la mujer estuviera fuera del rango de entre 14 y 42 años. Con esto se eliminaron 79 registros y se quedó una muestra de 9921 registros.

A esta muestra resultante se le eliminaron las combinaciones de edad y nivel de estudios imposibles, como por ejemplo, mujeres con edad de 14 años y más de 15 años de estudios. Para esto se asumió que la escolarización empezaba a los 6 años, por lo tanto, se eliminaron todos los registros donde la edad menos los años estudiados fuera menor que 6 años. De este modo, se eliminaron de una vez todos los registros inconsistentes.

La base de datos resultante quedó con 9158 registros, de esta se obtuvo una muestra sin reemplazamiento de 361 registros la cual sería la muestra final de mujeres.

Por último, se categorizaron las variables que se habían simulado de forma cuantitativa de la siguiente forma:

Edad:

- <20: mujeres entre 14 y 19 años
- [20-29]: mujeres entre 20 y 29 años
- ≥ 30: mujeres entre 30 y 42 años

Nivel de estudios:

- < Educación secundaria: mujeres con menos de 10 años estudiados
- Educación secundaria: mujeres entre 10 y 14 años estudiados
- Estudios universitarios: mujeres con más de 15 años estudiados

Para esta covariable se asumió que la educación secundaria empezaba después de 10 años de enseñanza obligatoria, es decir, a las mujeres con menos de 10 años de estudios se las categorizó como nivel de estudios menor que secundaria (17). Los estudios universitarios se plantearon en media con una duración de 4 años y que empezaban después de 12 años estudiados (17) (lo que equivaldría a después de terminar colegio e instituto), es decir, a las mujeres con al menos 16 años estudiados se las consideró con nivel de estudios universitarios.

A continuación se muestra un esquema con esta categorización:

Tabla 2 Categorización de la variable: Nivel de estudios

Número de años estudiados	Tipo de educación	Categoría
1	1º de educación obligatoria	< Educación secundaria
2	2º de educación obligatoria	< Educación secundaria
3	3º de educación obligatoria	< Educación secundaria
4	4º de educación obligatoria	< Educación secundaria
5	5º de educación obligatoria	< Educación secundaria
6	6º de educación obligatoria	< Educación secundaria
7	7º de educación obligatoria	< Educación secundaria
8	8º de educación obligatoria	< Educación secundaria
9	9º de educación obligatoria	< Educación secundaria
10	10º de educación obligatoria	< Educación secundaria
11	1º de educación secundaria	Educación secundaria
12	2º de educación secundaria	Educación secundaria
13	1º de universidad	Educación secundaria
14	2º de universidad	Educación secundaria
15	3º de universidad	Educación secundaria
≥16	4º de universidad	Estudios universitarios

Por último, para la variable estatus de tabaquismo previo los valores de 1 se categorizaron como < 20 cigarrillos diarios y los de 0 como ≥ 20 cigarrillos diarios.

4.1.4 Simulación de la variable respuesta

Para la simulación de la variable respuesta se utilizaron las modas de la sensibilidad y especificidad y los coeficientes estimados resumidos en la Tabla 1 de la siguiente forma: en la base de datos de las covariables se generaron 3 variables auxiliares adicionales, una por cada covariable. A estas se les dieron los valores de las estimaciones de los β correspondientes a cada categoría de la covariable de tal forma que por cada registro (mujer) el valor de estas variables auxiliares tenía el valor de los coeficientes estimados correspondientes a las categorías de sus covariables. Para las categorías de referencia el valor de las variables auxiliares fue 0. En la Tabla 3 se resume el valor de las variables auxiliares según las categorías de las covariables:

Tabla 3 Valores de las variables auxiliares

Covariable	Categoría	Variable auxiliar	Valor
Edad	<20	Aux1	0
	[20-29]		$\hat{\beta}_1 = -1.75$
	≥30		$\hat{\beta}_2 = -1.32$
Nivel de estudios	< Educación secundaria	Aux2	0
	Educación secundaria		$\hat{\beta}_3 = 0.86$
	Estudios universitarios		$\hat{\beta}_4 = 0.37$
Nivel de tabaquismo previo	<20 cigarrillos diarios	Aux3	$\hat{\beta}_5 = 1.29$
	≥ 20 cigarrillos diarios		0

A continuación, se calculó la probabilidad estimada asociada a la variable respuesta de cada registro de acuerdo a los valores de las variables auxiliares y el coeficiente estimado del término independiente mediante las fórmulas [1] y [2]. Una vez obtenidas estas se simuló la variable respuesta (0 o 1) aplicando la distribución bernoulli para cada registro asumiendo un ensayo con dichas probabilidades estimadas. Una vez hecho esto se obtuvo la base de datos simulada con las 3 covariables, edad, nivel de estudios y estatus de tabaquismo previo y la variable respuesta dejar de fumar, donde 0 corresponde a No y 1 a Sí.

4.1.5 Análisis descriptivo y comparación de resultados

En esta sección se presentan los resultados del análisis descriptivo de la base de datos simulada previamente. Una muestra de la base de datos simulada se puede ver en el anexo 7.1 y en la tabla siguiente (Tabla 4) se muestra el análisis descriptivo de las variables que la componen:

Tabla 4 Resumen descriptivo de las variables del estudio

Variable Categoría	Dejar de fumar (N=66)	No dejar de fumar (N=295)	Todas las mujeres (N=361)
Dejar de fumar [1]	66 (18.28%)	295 (81.72%)	361 (100%)
Edad:			
<20	9 (13,64%)	7 (2,37%)	16 (4.43%)
[20-29]	46 (69,7%)	241 (81,69%)	287 (79.50%)
≥30	11 (16,67%)	47 (15,93%)	58 (16.07%)
Nivel de estudios:			
<Educación secundaria	10 (15,15%)	65 (22,03%)	75 (20.78%)
Educación secundaria	54 (81,82%)	216 (73,22%)	270 (74.79%)
Estudios universitarios	2 (3,03%)	14 (4,75%)	16 (4.43%)
Estatus de tabaquismo previo:			
<20 cigarrillos diarios	60 (90,91%)	193 (65,42%)	253 (70.08%)
≥20 cigarrillos diarios	6 (9,09%)	102 (34,58%)	108 (29.92%)

Nota: Todos los porcentajes están calculados en base al total por columna excepto en [1] donde los porcentajes están basados en el total de mujeres.

Como se observa en la Tabla 4 se tiene una muestra de N=361 mujeres donde 66 (18. 28%) de ellas dejaron de fumar frente a 295 (81.72%) que no lograron dejar de fumar. En relación a las covariables, de manera global y, tanto en el grupo de mujeres que dejaron de fumar como en el grupo que no, se puede decir que las categorías de [20-29] años de edad, educación secundaria

como nivel de estudios y <20 cigarrillos diarios como estatus de tabaquismo previo fueron las más predominantes con porcentajes por encima del 65% en todas ellas. Por otro lado, se puede ver que había muy pocas mujeres con nivel de estudios universitarios 16 (4.43%) y muy pocas mujeres con edad <20 años 16 (4.43%).

Dado que hubo que depurar previamente la base de datos simulada eliminando los registros con las combinaciones entre edad y nivel de estudios imposibles, a continuación se muestra una tabla de contingencia de estas dos covariables:

Tabla 5 Tabla de contingencia entre edad y nivel de estudios

Edad/ Nivel de estudios	**<20 (N=16)**	**[20-29] (N=287)**	**≥30 (N=58)**
<Educación secundaria (N=75)	3	54	18
Educación secundaria (N=270)	13	220	37
Estudios universitarios (N=16)	0	13	3

Como se aprecia en la Tabla 5 no hay mujeres con niveles de estudios universitarios y edades menores de 20 años, lo cual se esperaba ya que es una de las combinaciones que no se podía dar.

Por último, se realizó el análisis comparativo de las covariables de la base de datos simulada y los datos de referencia obtenidos de los artículos. Como la base de datos simulada final contenía las covariables ya categorizadas, la comparación de las variables cuya información recogida en los artículos era de tipo cuantitativo se usaron las variables no categorizadas. En la Tabla 6 se muestran los resultados:

Tabla 6 Comparación resultados de la base de datos simulada y la información recogida en los artículos

Variable **Categoría**	**Resultados de los artículos**	**Resultados de la base de datos simulada**
Edad		
Media (desviación típica)	24.9 (4.7)	25.6 (3.9)
Nivel de estudios:		
Media (desviación típica)	12.3 (2.0)	12.2 (2.0)
Estatus de tabaquismo previo:		
<20 cigarrillos diarios, [n(%)]	254 (70.36%)	253 (70.08%)
≥20 cigarrillos diarios, [n(%)]	107 (29.64%)	108 (29.92%)

Como se aprecia para las variables edad y nivel de estudios la media y desviación típica entre ambas fuentes son prácticamente iguales del mismo modo que los porcentajes de las categorías de la variable estatus de tabaquismo. Sin embargo, las frecuencias absolutas de estas no lo son debido a que en el artículo de donde se extrajo esa información se habían analizado más mujeres.

4.2 ESTIMACIÓN DE LOS PARAMETROS DEL MODELO

4.2.1 Elicitación de los hiperparámetros de las distribuciones a priori

Como primer paso para la estimación de los parámetros del modelo $(Se, Sp, \boldsymbol{\beta})$ es necesario elicitar los hiperparámetros de las distribuciones a priori betas de la sensibilidad, especificidad y las CMPs.

La información a priori para la sensibilidad y especificidad fue basada en la opinión de un experto tal y como se comenta en el artículo de referencia (1). Al experto se le preguntó la probabilidad de que una mujer embaraza dijera que había dejado de fumar cuando realmente había dejado de fumar y la probabilidad de una mujer dijera que no había dejado de fumar cuando realmente no había dejado de fumar. La moda y los intervalos de probabilidad al 95% de la opinión del experto en relación a estas dos preguntas fueron usados para calcular los hiperparámetros de las betas para sensibilidad y especificidad siendo estos:

- Para la pregunta 1:

$$Moda = 1; IP_{95\%}(Se) = (0.96, 1.00) \rightarrow Beta_{Se}(99,1)$$

- Para la pregunta 2:

$$Moda = 0.93; IP_{95\%}(Sp) = (0.68, 0.98) \rightarrow Beta_{Sp}(14,2)$$

A continuación, en la Figura 1, se muestran las distribuciones a priori betas obtenidas para la sensibilidad y especificidad:

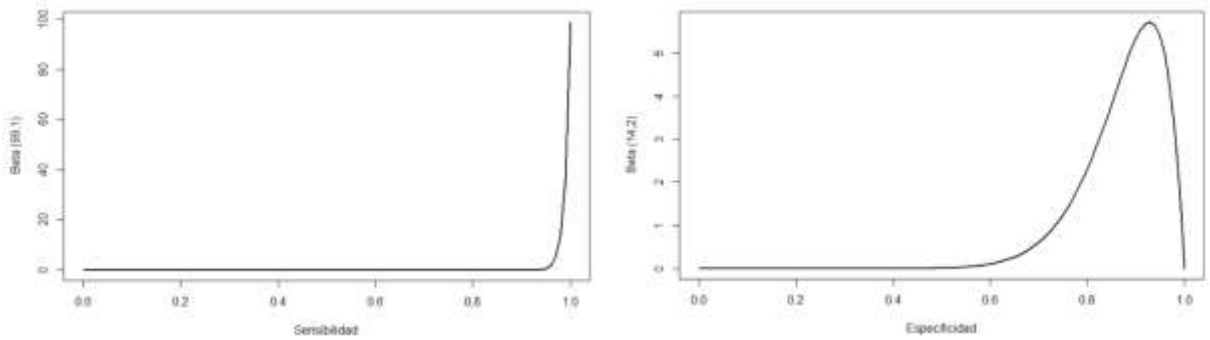

Figura 1 Distribuciones a priori beta de la sensibilidad y especificidad

Para la estimación de los coeficientes (**β**) se usarán las siguientes CMPs basadas en la opinión del experto sobre dejar de fumar en mujeres embarazadas para cada una de las combinaciones de covariables mostradas en la Tabla 7. El procedimiento para seleccionar las distribuciones a priori beta estaba basado en la moda y el intervalo de probabilidad al igual que para la sensibilidad y especificidad.

Tabla 7 CMPs seleccionadas en base a la opinión del experto para cada una de las combinaciones de covariables

	Especificación de la covariable $\tilde{x}_i$						Especificación a priori $\tilde{\pi}_i$	
i	Término independiente	Edad [20-29]	Edad ≥30	Nivel estudios Edc. secundaria	Nivel estudios Universitarios	Nivel de tabaquismo previo <20	$\tilde{\pi}_i \sim Beta(a_i, b_i)$	Moda a priori
1	1	1	0	1	0	1	$Beta(8,15)$	0.33
2	1	0	1	1	0	1	$Beta(10,15)$	0.39
3	1	1	0	1	0	0	$Beta(3,13)$	0.14
4	1	1	0	0	1	1	$Beta(8,10)$	0.44
5	1	1	0	0	0	1	$Beta(4,15)$	0.18
6	1	0	0	1	0	0	$Beta(6,15)$	0.26

4.2.2 Modelos

A continuación se describen los modelos propuestos en los objetivos de este TFM que se han analizado:

- Modelo 1: Regresión logística desde la perspectiva bayesiana asumiendo las distribuciones a priori para las CMPs, sensibilidad y especificidad descritas en la sección anterior.
- Modelo 2: Regresión logística desde la perspectiva bayesiana asumiendo las distribuciones a priori para la sensibilidad y especificidad descritas en la sección anterior y normales no informativas, $N(\mu = 0, \sigma = 0.001)$, para los parámetros de los coeficientes.

- Modelo 3: Regresión logística desde la perspectiva bayesiana asumiendo las distribuciones a priori para las CMPs descritas en la sección anterior y sensibilidad y especificidad iguales a 1.
- Modelo 4: Regresión logística desde la perspectiva bayesiana asumiendo las distribuciones a priori normales no informativas, $N(\mu = 0, \sigma = 0.001)$, para los parámetros de los coeficientes y sensibilidad y especificidad iguales a 1.
- Modelo 5: Regresión logística desde la perspectiva frecuentista asumiendo sensibilidad y especificidad iguales a 1.

Para todos los modelos las distribuciones a posteriori se han obtenido mediante los métodos MCMC utilizando la librería BRugs del programa R. Para todos se han usado 3 cadenas con al menos 30000 iteraciones y se han descartado al menos las 10000 primeras iteraciones.

4.2.2.1 Modelo 1

A continuación (Figura 2) se muestran los resultados del modelo 1:

```
> modeloBRugs2
$`Stats`
             mean      sd  MC_error  val2.5pc   median val97.5pc  start sample
OR_cigar  6.05900 2.81700 1.074e-02  2.659000  5.43600   13.1600 100001 120000
OR_edad1  0.17300 0.08348 3.408e-04  0.059880  0.15690    0.3810 100001 120000
OR_edad2  0.23120 0.12690 5.274e-04  0.070210  0.20390    0.5553 100001 120000
OR_educ1  2.12200 0.84450 3.486e-03  1.002000  1.95400    4.2190 100001 120000
OR_educ2  3.14500 1.80100 7.689e-03  0.988100  2.73300    7.7310 100001 120000
Se        0.98790 0.01188 4.658e-05  0.956000  0.99150    0.9997 100001 120000
Sp        0.97660 0.01455 6.126e-05  0.941700  0.97930    0.9967 100001 120000
beta1    -1.86200 0.46810 1.945e-03 -2.815000 -1.85200   -0.9649 100001 120000
beta2    -1.60000 0.52440 2.234e-03 -2.656000 -1.59000   -0.5883 100001 120000
beta3     0.68370 0.36430 1.570e-03  0.002089  0.66980    1.4400 100001 120000
beta4     1.00800 0.52300 2.344e-03 -0.011950  1.00600    2.0450 100001 120000
beta5     1.71500 0.40660 1.628e-03  0.978100  1.69300    2.5770 100001 120000
beta6    -1.66000 0.53080 2.450e-03 -2.727000 -1.65000   -0.6447 100001 120000
pi2[1]    0.24680 0.03404 1.393e-04  0.182400  0.24600    0.3157 100001 120000
pi2[2]    0.30020 0.05795 2.618e-04  0.192700  0.29820    0.4185 100001 120000
pi2[3]    0.05904 0.02165 9.105e-05  0.023690  0.05673    0.1074 100001 120000
pi2[4]    0.31610 0.08286 3.971e-04  0.166600  0.31200    0.4879 100001 120000
pi2[5]    0.14600 0.04181 1.865e-04  0.072660  0.14300    0.2357 100001 120000
pi2[6]    0.28060 0.07893 3.595e-04  0.142300  0.27500    0.4489 100001 120000

$DIC
       Dbar  Dhat   DIC    pD
y     317.8 314.4 321.2 3.401
total 317.8 314.4 321.2 3.401
```

Figura 2 Resultados de las distribuciones a posteriori de los parámetros del modelo 1

En la figura anterior se muestran los resultados de las medias, medianas, y varios percentiles de los parámetros estimados de la distribución a posteriori. Adicionalmente también se obtiene el valor del DIC que más adelante se usará para comparar los distintos modelos propuestos.

Por otro lado, la convergencia de las 3 cadenas empleadas para cada parámetro se comprueba mediante las gráficas de trazas y los criterios de Geweke y PSRF presentados en el anexo 7.2.1 (en la Figura 5, Figura 6 y Figura 7).

4.2.2.2 Modelo 2

Para el modelo 2 se procedió de igual manera que en modelo anterior pero no se consiguió alcanzar la convergencia de las cadenas por lo que los resultados no se mostrarán. La no convergencia de las cadenas se puede observar en el anexo 7.2.2 (en la Figura 8, Figura 9 y Figura 10)

4.2.2.3 Modelo 3, 4 y 5

Los resultados de los modelos 3, 4 y 5 se resumen en la sección siguiente con el fin de comparar los resultados obtenidos en todos ellos, ya que para estos tres modelos se asumen sensibilidad y especificidad iguales a 1.

Para los modelos 3 y 4 sí se alcanzó la convergencia de las cadenas, mediante los mismos procedimientos que los empleados para el modelo 1.

4.2.3 Comparación de resultados

Como último paso se comparan los resultados obtenidos con cada modelo, parámetros y OR estimados con cada uno de ellos. Esta información se resume en la tabla siguiente (Tabla 8 y Tabla 9):

Tabla 8 Comparación de los parámetros estimados de los modelo 1, 3, 4 y 5

Variable Categoría	Parámetro estimado	Modelo 1	Modelo 3	Modelo 4	Modelo 5
Término independiente	$\hat{\beta}_0$	-1.65* (-2.73, -0.65)	-1.58* (-2.58, -0.64)	-1.71* (-3.27, -0.24)	-1.60* (-3.08, -0.19)
Edad:					
<20	Referencia				
[20-29]	$\hat{\beta}_1$	-1.85* (-2.82, -0.97)	-1.72* (-2.56, -0.89)	-1.89* (-3.03, -0.79)	-1.86* (-2.99, -0.78)
≥30	$\hat{\beta}_2$	-1.59* (-2.69, -0.59)	-1.49* (-2.45, -0.55)	-1.76* (-3.06, -0.50)	-1.70* (-2.99, -0.48)
Nivel de estudios:					
<Educación secundaria	Referencia				
Educación secundaria	$\hat{\beta}_3$	0.67* (0.002, 1.44)	0.65* (0.03, 1.33)	0.63 (-0.12, 1.46)	0.59 (-0.14, 1,41)
Estudios universitarios	$\hat{\beta}_4$	1.01* (0.01, 2.05)	0.92 (-0.07, 1.87)	-0.06 (-2.19, 1.50)	0.06 (-1.91, 1.57)
Estatus de tabaquismo previo:					
≥20 cigarrillos diarios	Referencia				
<20 cigarrillos diarios	$\hat{\beta}_5$	1.70* (0.98, 2.58)	1.58* (0.94, 2.32)	1.75* (0.92, 2.79)	1.68* (0.87, 2.68)
	Se	0.992* (0.956, 0.999)	1	1	1
	Sp	0.979* (0.942, 0.997)	1	1	1

Nota: Para los modelo 1,3 y 4 se muestran las medianas y sus correspondientes intervalos de probabilidad (95%) de las distribuciones a posteriori de los coeficientes. Adicionalmente para el modelo 1 también se dan las medianas e IP (95%) de la sensibilidad y especificidad. Para el modelo 5 se muestran los coeficientes estimados y sus correspondientes intervalos de confianza (95%). Los coeficientes estimados cuyo IP o IC no contenga al 0 son destacados con un *.

Como se aprecia en la Tabla 8, para el modelo principal (modelo 1) todas las estimaciones de los coeficientes, así como, las estimaciones de la sensibilidad y especificidad se muestran relevantes,

dado que sus correspondientes IP o IC no contienen al 0. Sin embargo, en los demás modelos no sucede lo mismo. Para el modelo 3, la estimación del coeficiente asociado a los estudios universitarios no es relevante. En el modelo 4, las estimaciones de los coeficientes asociados a la variable nivel de estudios no son relevantes. Por último, el modelo frecuentista (modelo 5) presenta los mismos coeficientes significativos que el modelo 4.

Tabla 9 Comparación de los OR ajustados de los modelos 1, 3, 4 y 5

Variable Categoría	ORs ajustados	Modelo 1	Modelo 3	Modelo 4	Modelo 5
Edad:					
<20	Referencia				
[20-29]	$\widehat{OR}_1$	0.16* (0.06, 0.38)	0.18* (0.08, 0.41)	0.15* (0.05, 0.45)	0.15* (0.05, 0.45)
≥30	$\widehat{OR}_2$	0.20* (0.07, 0.56)	0.23* (0.09, 0.58)	0.17* (0.05, 0.61)	0.18* (0.05, 0.62)
Nivel de estudios:					
<Educación secundaria	Referencia				
Educación secundaria	$\widehat{OR}_3$	1.95* (1.00, 4.22)	1.91* (1.03, 3.77)	1.87 (0.89, 4.30)	1.81 (0.87, 4.10)
Estudios universitarios	$\widehat{OR}_4$	2.73* (0.99, 7.73)	2.50 (0.94, 6.59)	0.94 (0.11, 4.48)	1.06 (0.15, 4.81)
Estatus de tabaquismo previo:					
≥20 cigarrillos diarios	Referencia				
<20 cigarrillos diarios	$\widehat{OR}_5$	5.44* (2.66, 13.16)	4.87* (2.55, 10.16)	5.78* (2.51, 16.29)	5.38* (2.38, 14.52)

Nota: Para los modelo 1, 3 y 4 se muestran las medianas y sus correspondientes intervalos de probabilidad (95%) de las distribuciones a posteriori de los OR. Para el modelo 5 se muestran los OR ajustados y sus correspondientes intervalos de confianza (95%). Los ORs cuyo IP o IC no contiene al 1 son destacados con un *.

Dado que se trata de un modelo logístico, se presentan en la tabla anterior (Tabla 9) las estimaciones de los ORs dado que es importante su interpretación. Desde el punto de vista de qué ORs son relevantes en cada modelo, sucede lo mismo que con los coeficientes pero en este caso se determina que un OR es relevante cuando el IP o IC no contiene al 1. En el modelo 1, todos los

OR son relevantes, en el modelo 3 son todos relevantes menos el asociado a la categoría estudios universitarios de la variable nivel de estudios. En el modelo 4 y 5 sucede como con el modelo 3 añadiendo como no relevante el OR asociado a la categoría de educación secundaria de la variable nivel de estudios.

Desde el punto de vista de la interpretación de los ORs, en el modelo 1, los ORs que presentan una asociación positiva con dejar de fumar con respecto a la categoría de referencia son los correspondientes a las categorías: educación secundaria, estudios universitarios y menos de 20 cigarrillos diarios. Sin embargo, los ORs que presentan una asociación negativa con dejar fumar son las categorías de: edades entre 20-29 y mayores o iguales de 30. En los demás modelos, los ORs tienen la misma interpretación pero aquellos que no son relevantes no son interpretables dado que no se puede determinar si la asociación es positiva o negativa.

A continuación se muestran los DIC de los modelos bayesianos para el evaluar el ajuste global del modelo:

Tabla 10 DIC de los modelo 1, 3 y 4

Modelo	**DIC**
Modelo 1	321.2
Modelo 3	319.8
Modelo 4	321.7

Como se aprecia en la Tabla 10, el modelo 3 tiene el DIC más pequeño lo que indica que es el modelo que mejor se ajusta, siguiéndole el modelo 1 y luego el modelo 4.

En la Tabla 11 se muestra la comparación entre los resultados obtenidos del modelo 1 hecho en este TFM y los obtenidos en el artículo. Como se observa las estimaciones son similares excepto en el caso del coeficiente correspondiente a la categoría de estudios universitarios. Para este parámetro el modelo hecho en el TFM muestra una relación positiva y relevante y, sin embargo, en el modelo del artículo el valor de la mediana es bastante más pequeño y no muestra relevancia estadística. Por otro lado, cabe destacar la diferencia en los resultados de las estimaciones de la especificidad, en el modelo del TFM se muestran unas estimaciones mayores que en las del artículo.

Estas dos diferencias se pueden interpretar como que los datos simulados no parecen adoptar las mismas relaciones de los datos originales y, por lo tanto, las asunciones de las distribuciones a priori no funcionan del mismo modo que el modelo del artículo. Por este motivo pueden salir resultados diferentes para algunas estimaciones y se puede observar un ajuste del modelo menor del esperado.

Tabla 11 Parámetros estimados de los modelo 1 del TFM vs artículo

Parámetro estimado	Modelo 1	Modelo 1 del artículo
$\hat{\beta}_0$	-1.65* (-2.73, -0.65)	-1.27* (-2.27, -0.55)
$\hat{\beta}_1$	-1.85* (-2.82, -0.97)	-1.75* (-2.62, -1.14)
$\hat{\beta}_2$	-1.59* (-2.69, -0.59)	-1.32* (-2.22, -0.58)
$\hat{\beta}_3$	0.67* (0.002, 1.44)	0.86* (0.24, 1.64)
$\hat{\beta}_4$	1.01* (0.01, 2.05)	0.37 (-0.44, 1.28)
$\hat{\beta}_5$	1.70* (0.98, 2.58)	1.29* (0.68, 2.05)
Se	0.992* (0.956, 0.999)	0.992* (0.960,1.000)
Sp	0.979* (0.942, 0.997)	0.901* (0.799, 0.980)

5 CONCLUSIONES

En este trabajo se ha expuesto, gracias al enfoque bayesiano, un sencillo modelo logístico que incorpora sensibilidad y especificidad desconocidas. Adicionalmente, también, se ha podido incorporar diferente información a priori para la estimación de los coeficientes y comparar los resultados de los diferentes modelos realizados.

Mientras que con el modelo 1 se obtenían estimaciones de los parámetros y ORs relevantes, el resto de modelos se descartaba algún coeficiente dada su no relevancia.

Por otro lado, se ha demostrado que la información a priori incorporada en el modelo 1 sobre las combinaciones de covariables así como de la sensibilidad y especificidad proporciona un mejor ajuste (DIC=321.2) que el modelo 4 (DIC=321.7) el cual asumía sensibilidad y especificidad iguales a 1 y normales no informativas como a distribuciones a priori de los coeficientes del modelo. También destacar que el modelo 1 tiene mejor ajuste que el modelo 2, bajo el cual no se llegó a alcanzar la convergencia de las cadenas y, por lo tanto, no se pudieron estimar los parámetros del modelo.

Sin embargo, el modelo 3 proporcionó el mejor ajuste (DIC=319.8) de todos los realizados. Este es el que asumía sensibilidad y especificidad iguales a 1 y las mismas a priori para las combinaciones de covariables que el modelo 1. Esto lleva a la conclusión que la información a priori que se está considerando para el modelo 1 (y para el modelo 4) de la sensibilidad y especificidad no parece ser asumida por los datos. En concreto, la información a priori especificidad no parece ser seguida por las estimaciones obtenidas. Si se comparan las especificidades obtenidas en el modelo 1 del TMF y el artículo de referencia [$\widehat{Sp}_{TFM}\,(IP_{95\%}) = 0.979\ (0.942, 0.997)$ y $\widehat{Sp}_{\,artículo}(IP_{95\%}) = 0.901(0.799, 0.980)$] se observa que la obtenida en el artículo es menor y el límite inferior del intervalo de probabilidad es bastante menor que el obtenido en el modelo 1 del TFM ($0.942\ vs\ 0.799$).

Este problema puede ser causado por la base de datos simulada, a pesar de que se ha recogido información válida de los artículos de referencia y se ha incorporado para la simulación, no parece que se haya obtenido una base de datos exactamente con las mismas características. Uno de las variables que probablemente esté más afectada es el nivel de estudios dado que el coeficiente de la categoría de nivel de estudios universitarios es relevante para el modelo 1 del TMF mientras

que para el obtenido en el modelo del artículo no [$\hat{\beta}_{4_{TMF}} = 1.01(0.01, 2.05)$ y $\hat{\beta}_{4_{artículo}} = 0.37(-0.44, 1.28)$]. Otra posible causa es que exista alguna relación entre varias covariables que no han podido ser simuladas dado que no se disponía información de ello.

A pesar de esto último, se puede concluir que el enfoque bayesiano es una herramienta muy útil para poder incorporar información a priori acerca de los parámetros de un modelo.

6 BIBLIOGRAFÍA

1. McInturff P, Johnson WO, Cowling D, Gardner IA. Modelling risk when binary outcomes are subject to error. Statistics in Medicine. 2004; 23: p. 1095–1109.

2. Al Mansari A, Obeid Y, Islam N, Fariduddin M, Hassoun A, Djaballah K, et al. GOAL study: clinical and non-clinical predictive factors for achieving glycemic control in people with type 2 diabetes in real clinical practice. BMJ Open Diabetes Research and Care. 2018; 6(000519).

3. Sexton M, Hebel JR. A clinical trial of change in maternal smoking and its effect on birth weight. Journal of the American Medical Association. 1987; 251: p. 911-915.

4. Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technology Assessment. 2007; II(50).

5. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decision Making. 1991; 11: p. 88-94.

6. Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnosti c tests for ischemic heart disease. Journal of General Internal Medicine. 1988; 3: p. 476-481.

7. P D. Evaluating rapid tests for streptococcal pharyngitis: the apparent accuracy of a diagnostic test when there are errors in the standard of comparison. Medical Decision Making. 1987; 7: p. 92-96.

8. MH Z. Evaluation of the clinical accuracy of laboratory tests. Archives of Pathology & Laboratory Medicine. 1988; 112(4): p. 383-386.

9. Rosner B. Fundamentals of Biostatistics. 7th ed. Brooks/Cole CL, editor. Boston; 2011.

10. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. American Journal of Epidemiology. 1997; 146: p. 195-203.

11. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Online].; 2008. Disponible en: https://www.R-project.org/.

12. J. Dobson A. An introduction to Generalized linear models. 2nd ed. CHAPMAN H, editor. Florida; 2002.

13. Silva L.C, Benavides A. El enfoque bayesiano: otra manera de inferir. Gac Sanit. 2001; 15(4): p. 341-346.

14. Gelman A, B. Carlin J, S. Stern H, B. Dunson D, Vehtari A, B. Rubin D. Bayesian Data Analysis. 3rd ed. HALL/CRC C, editor. New York; 2013.

15. J. Bedrick E, Christensen R, Johnson W. A New Perspective on Priors for Generalized Linear Models. Journal of the American Statistical Association. 1996 Dec: p. 1450-1460.

16. Thomas A, O'Hara B, Ligges U, Sturtz S. Making BUGS Open. R News. 2006; 6(1): p. 12-17.

17. Statistics UIo. UNESCO UIS. [Online].; 2021 [citado 05 Mayo 2021]. Disponible en: http://uis.unesco.org/en/country/us.

7 ANEXOS

7.1 ANEXO 1: BASE DE DATOS SIMULADA

ID	AGE	EDUCATION	SMOKING
1	20-29	<E.secundaria	>=20
2	20-29	E.secundaria	>=20
3	20-29	E.secundaria	<20
4	>=30	E.secundaria	<20
5	20-29	E.secundaria	<20
6	20-29	E.secundaria	>=20
7	20-29	<E.secundaria	<20
8	20-29	E.secundaria	<20
9	20-29	E.secundaria	>=20
10	20-29	E.secundaria	>=20
11	20-29	E.secundaria	>=20
12	>=30	E.secundaria	>=20
13	20-29	E.secundaria	<20
14	20-29	E.secundaria	>=20
15	20-29	E.secundaria	<20
16	20-29	<E.secundaria	>=20
17	20-29	E.secundaria	<20

Figura 3 Extracto de la base de datos simulada

7.2 ANEXO 2: CONVERGENCIA DE LAS CADENAS

7.2.1 Modelo 1

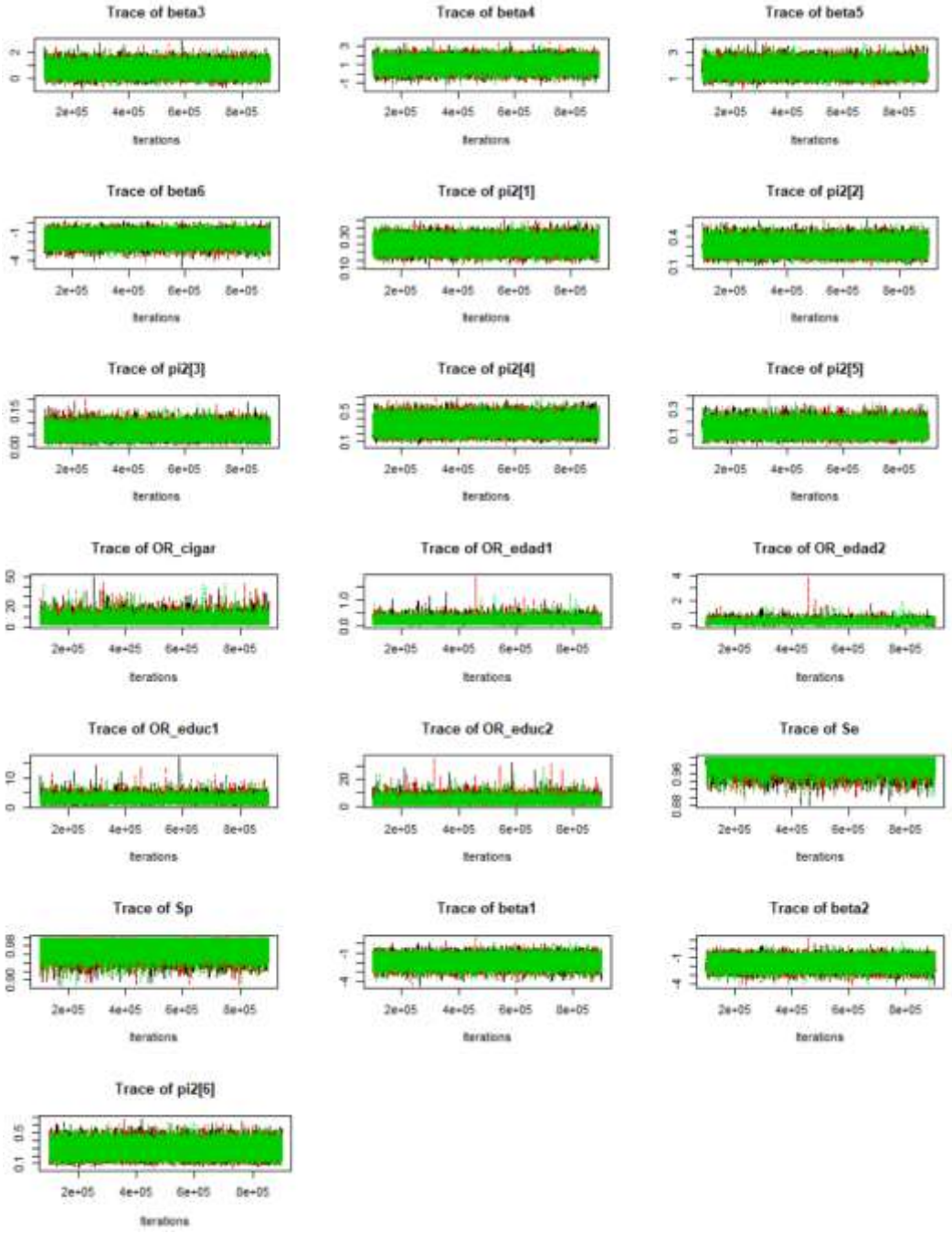

Figura 5 Gráficas de rachas correspondientes a las 3 cadenas del modelo 1

```
Potential scale reduction factors:

          Point est. Upper C.I.
OR_cigar           1          1
OR_edad1           1          1
OR_edad2           1          1
OR_educ1           1          1
OR_educ2           1          1
Se                 1          1
Sp                 1          1
beta1              1          1
beta2              1          1
beta3              1          1
beta4              1          1
beta5              1          1
beta6              1          1
pi2[1]             1          1
pi2[2]             1          1
pi2[3]             1          1
pi2[4]             1          1
pi2[5]             1          1
pi2[6]             1          1

Multivariate psrf

1
```

Figura 6 Criterio de PSRF del modelo 1

```
[[1]]

Fraction in 1st window = 0.2
Fraction in 2nd window = 0.4

OR_cigar OR_edad1 OR_edad2 OR_educ1 OR_educ2       Se       Sp    beta1    beta2    beta3   beta4
  0.3606   0.1259  -0.2308  -1.4056  -1.3912  -0.6577  -0.5297   0.3589  -0.1117  -1.4721  -1.388
   beta5    beta6   pi2[1]   pi2[2]   pi2[3]   pi2[4]   pi2[5]   pi2[6]
  0.1941   0.1221  -1.3392  -1.5562  -0.5561  -1.0789   0.9123  -1.2378


[[2]]

Fraction in 1st window = 0.2
Fraction in 2nd window = 0.4

OR_cigar OR_edad1 OR_edad2 OR_educ1 OR_educ2       Se       Sp    beta1    beta2    beta3   beta4
 0.09782  1.68907  1.25193  0.15984  0.83056  0.62286 -1.12917  1.33212  1.27951  0.29399  1.2671
   beta5    beta6   pi2[1]   pi2[2]   pi2[3]   pi2[4]   pi2[5]   pi2[6]
-0.17841 -0.97725  0.53681  0.59734  0.84788  1.46372 -0.06878 -0.86508


[[3]]

Fraction in 1st window = 0.2
Fraction in 2nd window = 0.4

OR_cigar OR_edad1 OR_edad2 OR_educ1 OR_educ2       Se       Sp    beta1    beta2    beta3   beta4
 1.11587 -0.01442  0.06653  0.93055  0.01697 -0.89342 -0.44850  0.16129  0.13167  1.03215  0.4546
   beta5    beta6   pi2[1]   pi2[2]   pi2[3]   pi2[4]   pi2[5]   pi2[6]
 1.33936 -1.55211  0.74016  0.47140 -1.19306 -0.14290 -0.80485 -1.04822
```

Figura 7 Criterio Geweke del modelo 1

7.2.2 Modelo 2

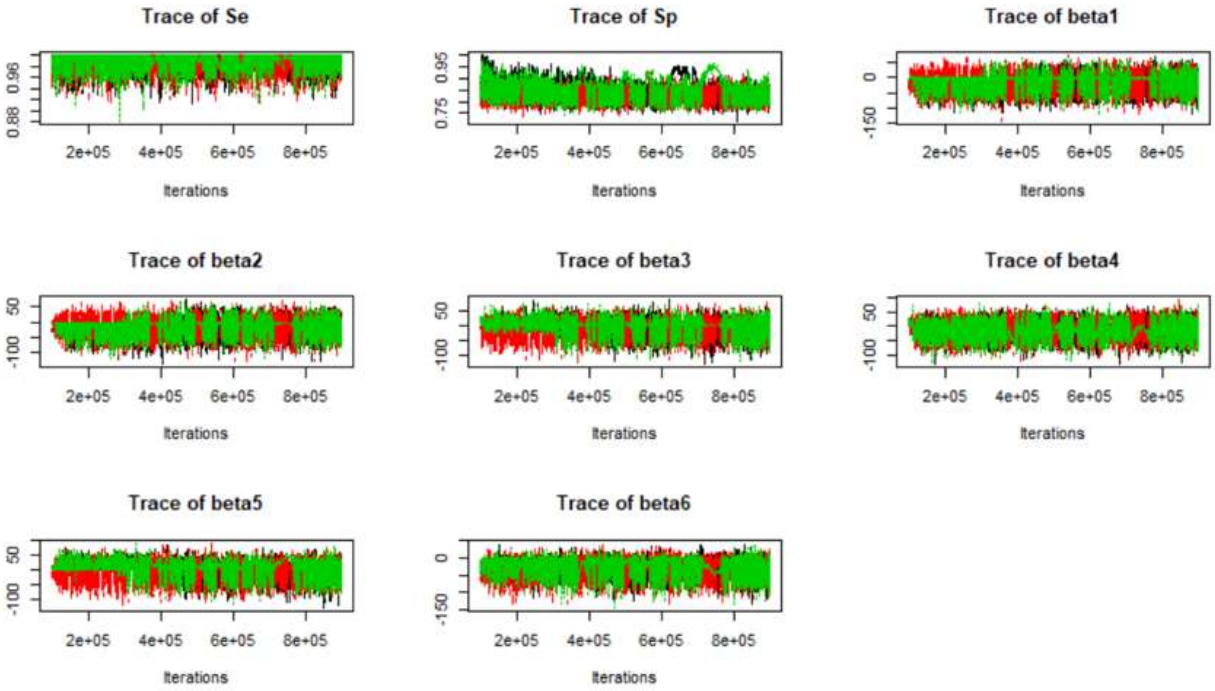

Figura 8 Gráficas de rachas correspondientes a las 3 cadenas del modelo 2

```
Potential scale reduction factors:

      Point est. Upper C.I.
Se          1.00       1.00
Sp          1.08       1.26
beta1       1.01       1.03
beta2       1.00       1.01
beta3       1.01       1.03
beta4       1.00       1.01
beta5       1.02       1.07
beta6       1.01       1.02

Multivariate psrf

1.06
```

Figura 9 Criterio de PSRF del modelo 2

```
[[1]]

Fraction in 1st window = 0.2
Fraction in 2nd window = 0.3

     Se      Sp   beta1   beta2   beta3   beta4   beta5   beta6
-0.2979  4.3805  2.8182  1.9710  6.0939 -0.2518  5.4573  5.9372


[[2]]

Fraction in 1st window = 0.2
Fraction in 2nd window = 0.3

     Se      Sp   beta1   beta2   beta3   beta4   beta5   beta6
 1.6399  0.9107  2.1786  0.7479 -0.2399 -1.6930  0.4360  1.9143


[[3]]

Fraction in 1st window = 0.2
Fraction in 2nd window = 0.3

     Se      Sp   beta1   beta2   beta3   beta4   beta5   beta6
-0.4352  0.8514  0.2626  0.2978  5.6256 -0.7124  2.7354  5.0717
```

Figura 10 Criterio Geweke del modelo 2