Análisis de la Prevalencia de Pre-diabetes y Diabetes Mellitus mediante Modelos STAR Multinomiales

Carlos Matías Hisgen[1]

Trabajo de fin de Master

Programa Oficial de Posgrado "Máster en Técnicas Estadísticas"

Ciclo 2007-2009

Universidad de Santiago de Compostela

Directores: Carmen Cadarso Suárez y César Sánchez Sellero

Junio de 2009

[1]Becario de la Agencia Española de Cooperación Internacional para el Desarrollo (AECID).

Los Profesores de universidad Carmen María Cadarso Suárez y César Andrés Sánchez Sellero, autorizan la presentación del presente Trabajo Fin de Master titulado "Análisis de la Prevalencia de Pre-diabetes y Diabetes Mellitus mediante Modelos STAR Multinomiales", realizado por el alumno Cárlos Matías Hisgen y del cual han sido directores.

Santiago de Compostela, 24 de Junio de 2009.

Firman en conformidad:

Carmen Cadarso Suárez ..... , César Sánchez Sellero .....

Índice

Agradecimientos

1. Introducción

2. Modelo de regresión lineal con respuesta multinomial
2.1 Modelo lineal generalizado (GLM) 7
2.2 Modelo GLM con respuesta binaria 9
2.3 Modelo de regresión multinomial 10

3. Modelo Aditivo Estructurado Generalizado (STAR)
3.1 El Modelo Aditivo Generalizado (GAM) basado en P-splines 13
3.2 El Modelo STAR 17
3.3. Inferencia con el STAR multinomial 20

4. Explorando la Prevalencia de Pre-diabetes y Diabetes
4.1 Origen y descripción de los datos utilizados 25
4.2 Prevalencia de diabetes y pre-diabetes según sexo, edad, IMC y OC 27
4.3 Evaluando factores de riesgo para la DM y PreDM 33

5. Modelos con respuesta nominal y ordinal en el análisis de la prevalencia
5.1 IMC y OC como factores de riesgo para la diabetes y pre-diabetes 36
5.2 Modelo Logit Multinomial Acumulativo para explicar la diabetes y pre-diabetes.. 43
5.3 Comparativa de modelos mediante curvas ROC 46

6. Conclusiones
6.1 Conclusiones desde el punto de vista biomédico 52
6.2 Conclusiones desde el punto de vista estadístico 53

Referencias bibliográficas 55

Agradecimientos

Adentrarse en el mundo de la investigación bioestadística, poseyendo una formación de economista, no es siempre una tarea sencilla. Por ello agradezco a mis directores, Carmen y César, por la ayuda y el aliento que me han sabido brindar. También estoy agradecido con María Xosé Rodriguez Álvarez, por la ayuda recibida en algunos temas computacionales.

1. Introducción

La *Diabetes Mellitus*, también conocida como *Diabetes de tipo 2*, es reconocida como un problema de salud de magnitud internacional. De hecho, estimaciones recientes proyectan un número elevado de pacientes diabéticos, por ejemplo, 300 millones para el año 2025 (King and Aubert, 1998) o 353 millones en 2030 (Yach et. al., 2006). Las personas diabéticas están expuestas a un mayor riesgo de padecer complicaciones de la salud, como pueden ser paros cardíacos, neuropatías, retinopatías y nefropatías. Estos trastornos vinculados a la diabetes pueden ser prevenidos, por lo que resulta de gran relevancia el diagnóstico temprano y la realización de campañas de prevención. Por este motivo, tienen mucha importancia los estudios epidemiológicos encaminados a refinar los métodos de diagnóstico, tanto de la diabetes como de estados pre-diabéticos.[2]

En la ejecución de los mencionados estudios, una herramienta frecuentemente utilizada es el análisis de regresión. En el contexto que nos concierne, los modelos de regresión son empleados para relacionar a la probabilidad de padecer un estado diabético (o pre-diabético) con un conjunto de factores de riesgo. De tales factores los mas importantes e interesantes de analizar son el sobrepeso y la obesidad.[3] Estos factores suelen ser medidos a través de índices o cantidades que representan variables continuas, desde el punto de vista estadístico. Para modelar el efecto de estas variables es común el uso de especificaciones paramétricas de la ecuación de regresión, siendo lo más típico el empleo del Modelo Lineal Generalizado (GLM).

Si bien los GLM demuestran ser de utilidad en un gran número de aplicaciones[4], existen contextos en los que su forma funcional paramétrica resulta ser demasiado rígida para describir adecuadamente el efecto de algún factor de riesgo específico. Es en esta situación donde los Modelos Aditivos Generalizados (GAM) son los más idóneos para el estudio del fenómeno.

En el presente trabajo se emplea una reciente clase de modelos denominados STAR (Generalized Structured Additive Regression), la cual incluye a los GAM y la extensión de éstos al caso de respuesta multinomial. El objetivo general del trabajo es ilustrar cómo los modelos STAR pueden ser usados para derivar información útil en el refinamiento de los diagnósticos médicos de la diabetes y prediabetes. Para ello se evalúan diversos modelos alternativos que explican la prevalencia de pre-diabetes y de diabetes utilizando una muestra de personas representativas de la población gallega.

[2]Existen trabajos recientes en esta área, como por ejemplo Anjana et. al. (2004), Jean-Baptiste et. al. (2006) y Faeh et. al. (2007).

[3]Para conocer la magnitud del problema de la obesidad y el sobrepeso en Galicia, puede consultarse Botana et. al. (2007).

[4]Véase por ejemplo Kim et. al. (2006)

En los Capítulos 2 y 3 se describen los métodos empleados partiendo de los modelos paramétricos más clásicos hasta llegar a los modelos STAR, todo bajo una notación común. Tal descripción representa una aportación adicional del trabajo.

El Capítulo 2 introduce el modelo lineal de respuesta multinomial, partiendo del Modelo Lineal Clásico de regresión y pasando por el GLM. En el Capítulo 3 se describe el modelo GAM basado en P-splines, representándolo luego como un modelo STAR. Las aplicaciones al problema médico bajo estudio se exponen en los capítulos 4 y 5. Finalmente, el Capítulo 6 resume las principales conclusiones.

2. Modelo de regresión lineal con respuesta multinomial

En este capítulo se describe el modelo de regresión lineal multinomial como una extensión del Modelo Lineal Generalizado (GLM) de respuesta univariada. El modelo GLM univariado puede extenderse al contexto de respuesta multivariante, dentro del cual el modelo multinomial es un caso particular (véase Fahrmeir and Tutz, 1994, cap. 3).

2.1 Modelo lineal generalizado (GLM)

Partiendo del modelo de regresión lineal múltiple con regresores estocásticos, y definiendo con y a la variable respuesta (variable explicada) y con $x = (x_1, \ldots, x_m)$ al vector de covariables (variables explicativas o regresores), se tiene la siguiente función de regresión que relaciona a y con el vector x

$$y = u'\beta + \epsilon, \tag{1}$$

en donde $u' = u(x)$ es el vector de diseño, el cual consiste en una función apropiada de x definida por el investigador, β es un vector de parámetros desconocidos y ϵ es el error aleatorio.

Para un conjunto de datos, compuesto por observaciones independientes e idénticamente distribuidas, de tipo corte transversal (y_i, x_i) con $i = 1, \ldots, n$, el modelo (1) implica que cada observación puede definirse como

$$y_i = u_i'\beta + \epsilon_i. \tag{2}$$

Al suponer que ϵ_i es i.i.d. y posee una distribución normal, estamos en el ámbito del Modelo Lineal Clásico, en el cual se verifica

$$y_i \sim N\left(\mu_i(x_i), \sigma^2\right) \;\; i = 1, \ldots, n,$$

siendo $\mu_i(x_i) = E(y_i|x_i) = u_i'\beta$ la esperanza de la variable respuesta y_i condicionada a las covariables x_i.

Expresando de esta manera al Modelo Clásico, es sencillo definir el modelo GLM relajando convenientemente dos de los supuestos recién expuestos. Uno de estos supuestos es el de la distribución de y_i condicionada a x_i, el cual puede ser ampliado a un conjunto de distribuciones posibles. En esta línea, podemos establecer un primer supuesto específico del modelo GLM:

- *Supuesto distributivo*: dado x_i, la distribución de y_i pertenece a la familia exponencial com media $\mu_i(x_i) = E(y_i|x_i)$ y un posible parámetro de escala ϕ que no depende de i.

El otro supuesto a ser relajado es el de la forma en que se relaciona la esperanza condicional $E(y_i|x_i)$ con el predictor lineal $\eta_i = u_i'\beta$. En el modelo clásico dicha relación está definida por la función identidad. En el modelo GLM las posibilidades para esta relación se amplían, quedando reflejadas en el segundo supuesto diferencial del modelo GLM:

- *Supuesto estructural*: la esperanza condicional $\mu_i(x_i) = E(y_i|x_i)$ se relaciona con el predictor lineal $\eta_i = u_i'\beta$ de la siguiente manera

$$\mu_i(x_i) = h(\eta_i) = h(u_i'\beta),$$

por lo que también es posible definir

$$\eta_i = g\left(\mu_i(x_i)\right),$$

en donde h es una función suficientemente suave, denominada *función respuesta* y $g = h^{-1}$ es la denominada *función link*.

En resumen, el modelo GLM queda totalmente caracterizado mediante las siguientes tres componentes:

- Una *componente aleatoria*, que especifica la distribución de probabilidades (condicionada en x) de la variable respuesta.

- La *componente sistemática* o predictor lineal (η), que consiste en una función lineal del vector de diseño $u(x)$.

- La *componente de unión* o función link, la cual define la forma funcional que relaciona a la componente sistemática con la esperanza condicional de la componente aleatoria ($E(y|x)$).

En lo que respecta a la componente aleatoria, ésta es susceptible de ser representada en forma general mediante la función de densidad (condicional en x) de la variable y. Dicha densidad suele ser presentada como

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi))\right\}, \tag{3}$$

en donde

$\theta_i(\mu_i)$ es el denominado *parámetro natural*, siendo una función dependiente de μ_i,

ϕ es un parámetro de dispersión o escala, y

$b(\cdot)$ y $c(\cdot)$ son funciones específicas de cada distribución en cuestión.

La forma específica que toman los parámetros θ_i y ϕ junto con las funciones $b(\cdot)$ y $c(\cdot)$ depende de la distribución específica que se considere.

En relación a la componente de unión, existe una función link (g) *natural* o *canónica* para cada distribución particular de la familia exponencial. Dicha función link relaciona directamente al parámetro natural con el predictor lineal, es decir:

$$\theta(\mu) = h^{-1}(\mu) = g(\mu) = \eta = u'\beta$$

2.2 Modelo GLM con respuesta binaria

Cuando la variable respuesta y es binaria, es decir $y \in (0,1)$, la distribución que le corresponde es la de Bernoulli, que es un caso particular de la Binomial. Entonces se tiene

$$y = \begin{cases} 1 \text{ con probabilidad } \pi, \\ 0 \text{ con probabilidad } (1-\pi), \end{cases}$$

por lo tanto $y \sim B(1, \pi)$.

Para el contexto de la regresión, lo relevante es la esperanza condicional de y, que en este caso es igual a la probabilidad de que $y = 1$, es decir

$$E(y|x) = P(y = 1|x) = \pi(x),$$

implicando a su vez la siguiente estructura de varianza condicional

$$var(y|x) = \pi(x)[1 - \pi(x)],$$

en donde $\pi(x)$ denota la dependencia de la probabilidad π respecto del vector de covariables x. Escogiendo la función link canónica se obtiene el modelo Logit, para el cual la función de regresión viene dada por

$$g(\pi) = \log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \eta = u'\beta \tag{4}$$

lo que implica una distribución logística como función respuesta

$$h(\eta) = \pi(x) = \frac{\exp(\eta)}{1+\exp(\eta)}. \tag{5}$$

La expresión (5) se puede escribir para un conjunto de n observaciones de la siguiente forma

$$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}, \ i = 1, \ldots, n. \tag{6}$$

Si bien existen otras funciones link posibles, es el modelo Logit el que es utilizado en este trabajo.

2.3 Modelo de regresión multinomial

A menudo surgen problemas en los cuales es necesario modelar un fenómeno mediante una variable categórica y relacionarla con un vector de covariables a través de un modelo de regresión. En tal caso la variable respuesta Y toma valores en un conjunto de k categorías, $Y \in \{1, \ldots, k\}$. Alternativamente, Y se puede representar como una variable respuesta multivariante, esto queda mas claro al definirla como un *vector* de variables binarias $y = (y_1, \ldots, y_q)$ con $q = k - 1$, siendo entonces sus q componentes

$$y_r = \begin{cases} 1, \text{ si } Y = r, \ r = 1, \ldots, q \\ 0, \text{ en otro caso,} \end{cases}$$

en donde $q = k - 1$, dado que se ha prescindido de la última categoría k. Como se verá mas adelante, en el marco del modelo de regresión, a k se la denomina como *categoría de referencia.*

Por consiguiente, la probabilidad de ocurrencia de cada una de las $k - 1$ categorías se puede expresar como

$$P(Y = r) = P(y_r = 1) \ r = 1, \ldots, q$$

Entonces, para definir el modelo de regresión con vector respuesta y, se extienden al caso multivariante los conceptos del modelo de respuesta binaria vistos en la sección previa. En este sentido se puede definir el vector de probabilidades $\pi_i = \mu_i = E(y_i|x_i)$, siendo $\pi_i = (\pi_{i1}, \ldots, \pi_{iq})$ un vector de dimensión $(q \times 1)$. El modelo de regresión multivariante viene dado por

$$\pi_i = (h(u_i'\beta_1), \ldots, h(u_i'\beta_q)), \tag{7}$$

en donde u_i' es el vector de diseño, β_r es el vector de parámetros desconocidos de la categoría r y h es la función respuesta . De manera equivalente a (7), el modelo puede escribirse como

$$g(\pi_i) = (g(\pi_{i1}), \ldots, g(\pi_{iq})) = (u_i'\beta_1, \ldots, u_i'\beta_q), \tag{8}$$

siendo g la función link. Al igual que en el caso binario, distintas funciones link darán origen a diferentes modelos.

Es muy importante aclarar que esta extensión, desde el modelo univariado binario al caso multivariante multinomial, supone que la variable categórica de respuesta es de tipo *nominal*, es decir, no existe un orden o jerarquía alguna entre las categorías representadas.

Entonces, en el contexto de respuesta categórica nominal y función link canónica, la cual implica una función respuesta como la definida en (5), se tiene el modelo Logit Multinomial. Para este caso, la probabilidad de pertenencia a la categoría r viene dada por

$$\pi_{ir} = P(Y_i = r|x_i) = \frac{\exp(u_i'\beta_r)}{1 + \Sigma_{s=1}^{q} \exp(u_i'\beta_s)},\ r = 1, \ldots, q. \tag{9}$$

La formulación (9) torna un tanto difícil la interpretación de los parámetros β en términos de su impacto sobre la probabilidad, por ello suele ser útil la representación del modelo a partir de la expresión (8), mediante el uso de la función link dada en (4). De esta manera surge la siguiente formulación:

$$g(\pi_{ir}), = \log\left\{\frac{\pi_{ir}}{1 - (\pi_{i1} + \ldots + \pi_{iq})}\right\} = u_i'\beta_r, \tag{10}$$

en donde el denominador dentro del logaritmo es la probabilidad de la categoría de referencia k. Por ello, la expresión (10) se puede reescribir como la ecuación

$$\log\left\{\frac{P(Y_i = r|x_i)}{P(Y_i = k|x_i)}\right\} = u_i'\beta_r, \tag{11}$$

la cual expresa al logaritmo del cociente de probabilidades como una función lineal del vector de diseño. Como puede apreciarse, siempre es importante definir convenientemente la categoría de referencia k, ya que en relación a ella se construyen los ratios de probabilidades y se interpretan los resultados que surgen al estimar el modelo.

Por otro lado, cuando se modela una variable respuesta categórica *ordinal*, resulta lógico explotar el ordenamiento de las k categorías. En esta situación carece de sentido el seleccionar una categoría fija de referencia, de entre las k categorías disponibles, ya que tal categoría no mantendría la misma "distancia ordinal" respecto de las demás.

Un posible enfoque es el de reducir las k categorías a solo dos categorías y modelar el fenómeno como en el GLM de respuesta binaria visto en la Sección 2.2. Dada una "categoría límite" r, la idea consiste en agrupar las categorías superiores a r (según su ordenamiento), por un lado, y agrupar las categorías inferiores a r (junto con r), por el otro. Cabe notar que estas dos nuevas "categorías acumuladas", cambian su composición según varía la "categoría límite" r que se tome.

Una vez definidas las dos "categorías acumuladas" el objetivo es modelar el siguiente ratio de probabilidades entre las mismas

$$\frac{P(Y \leq r|x)}{P(Y > r|x)} = \frac{\pi_1 + \ldots + \pi_r}{\pi_{r+1} + \ldots + \pi_k},\ r = 1, \ldots, q,$$

con $q = k - 1$. A diferencia del modelo con respuesta nominal, aquí no se toma la última categoría k como la de referencia, no obstante r no toma el valor k porque de hacerlo la expresión anterior estaría indefinida, ya que $P(Y > r = k|x)$ es igual a cero.

Equivalentemente, podemos reescribir este ratio de probabilidades como

$$\frac{P(Y \leq r|x)}{1 - P(Y \leq r|x)} = \frac{\pi_1 + \ldots + \pi_r}{1 - (\pi_1 + \ldots + \pi_r)} = \frac{\Pi^{(r)}}{1 - \Pi^{(r)}}, \ r = 1, \ldots, q,$$

y aplicando la función link canónica logística (de forma análoga a lo hecho en (4) para el modelo binario), se obtiene el modelo de regresión Logit Multinomial Acumulativo

$$\log \left\{ \frac{\Pi_i^{(r)}}{1 - \Pi_i^{(r)}} \right\} = \log \left\{ \frac{P(Y_i \leq r|x_i)}{P(Y_i > r|x_i)} \right\} = \beta_0^{(r)} + u_i'\beta. \tag{12}$$

Como siempre, es posible definir el mismo modelo utilizando la función respuesta h, que en el modelo Logit está dada por (5), de este modo se tiene

$$P(Y_i \leq r|x_i) = \Pi_i^{(r)} = h(\beta_0^{(r)} + u_i'\beta) = \frac{\exp(\beta_0^{(r)} + u_i'\beta)}{1 + \exp(\beta_0^{(r)} + u_i'\beta)}. \tag{13}$$

A diferencia del modelo Logit Multinomial con respuesta nominal dado por (11) y (9), en el presente modelo el vector de parámetros β no depende de r, es decir, es el mismo para todas las categorías $r = 1, \ldots, k$. El único parámetro que cambia es el intercepto $\beta_0^{(r)}$ (el cual no es determinante del efecto de las covariables x). Esta es la razón por la que se ha hecho explícita la inclusión del intercepto.

Por este motivo el modelo dado por (12) y (13) es llamado *modelo de odds proporcionales*, ya que verifica la siguiente igualdad

$$\log \left\{ \frac{P(Y \leq r|u^1)/P(Y > r|u^1)}{P(Y \leq r|u^2)/P(Y > r|u^2)} \right\} = (u^1 - u^2)'\beta,$$

en donde $u^1 \neq u^2$ son vectores de diseño correspondientes a dos poblaciones (o individuos) diferentes. Esto significa que el logaritmo de la razón de los *odds acumulados* es proporcional a la distancia entre los vectores u^1 y u^2, y no depende de la categoría r.

Como es habitual en el caso paramétrico, los modelos presentados se estiman por máxima verosimilitud. Los métodos de estimación se presentan con mas detalle para los modelos más flexibles que se exponen en el siguiente capítulo.

3. Modelo Aditivo Estructurado Generalizado (STAR)

En el Capítulo 2 se ha descrito el modelo de regresión multinomial, dentro del contexto del Modelo Lineal Generalizado (GLM). En este capítulo, dentro de la Sección 3.1, se extiende el GLM hacia su versión mas flexible, el Modelo Aditivo Generalizado (GAM) basado en P-splines.

Luego, en la Sección 3.2 se describe el Modelo Aditivo Estructurado Generalizado de Regresión (STAR), que plantea una interpretación alternativa del GAM para fines de estimación. El modelo STAR puede ser estimado mediante una metodología general conocida como *Bayesian P-splines*, la cual puede ser implementada mediante dos métodos. El primero es conocido como *Empirical Bayes estimation* (EB), que es la empleada en este trabajo. Existe otro método de estimación denominado *Full Bayesian estimation* basado en simulación MCMC (*Markov Chain Monte Carlo*) que no es considerado aquí.[5]

Finalmente, en la Sección 3.3, se expone la metodología EB de inferencia con los modelos STAR multinomiales.

3.1 El Modelo Aditivo Generalizado (GAM) basado en P-splines

Siguiendo la lógica del Capítulo 2 (Sección 2.1) en esta sección se extiende el Modelo Lineal Clásico hacia una versión más flexible, el Modelo Aditivo (MA) de regresión, que viene dado por

$$y_i = v_i'\gamma + \sum_{j=1}^{l} f_j(x_{ji}) + \epsilon_i,\ i = 1, \ldots, n. \tag{14}$$

Como antes, la variable y_i es la respuesta, v_i es un vector de diseño cuyas componentes se relacionan linealmente con la respuesta, γ es un vector de parámetros desconocidos, los x_j son elementos del vector $x_i = (x_{1i}, \ldots, x_{li})$ de variables continuas y ϵ_i es el error aleatorio i.i.d. que puede, o no, estar normalmente distribuido. Lo que distingue al MA respecto del Modelo Clásico son las l componentes aditivas $f_j(x_j)$, las cuales representan funciones suaves de las covariables continuas x_j. Esta innovación permite modelar no linealmente el efecto de un conjunto de covariables continuas, cuyo efecto (no lineal) no sea suceptible de ser modelado paramétricamente.

El problema que surge en esta instancia es el de cómo representar los términos f_j para poder luego estimarlos con los datos. Una forma de hacerlo es definiendo un espacio de funciones para cada uno de los f_j, que contenga al propio término f_j o, al menos, una buena

[5]Para consultar una descripción del método Full Bayesian véase Lang and Brezger, (2004).

aproximación del mismo. Para definir dicho espacio es necesario elegir una *base*, es decir, un conjunto de *funciones base*.

Una posibilidad es utilizar una base de splines cúbicos, aunque existen varias formas de representarla. En términos generales, un spline cúbico es una curva construida en mediante la unión de secciones de polinomios de grado 3. Dichos polinomios se unen de tal forma que la curva resultante es continua, al igual que su primera y segunda derivada. Los puntos en los cuales se unen las secciones polinómicas se denominan nodos (*knots*) y deben ser colocados a lo largo del rango de x_j. Dados t_j nodos colocados, en forma equidistante unos de otros, dentro el rango de x_j se tiene

$$\zeta_{j0} = x_{j,min} < \zeta_{j1} <, \ldots <, \zeta_{j,t_j-1} < \zeta_{j,t_j} = x_{j,max}.$$

Utilizando una base de splines de regresión cúbicos (*cubic regression splines*), definida por una base de B-splines de orden 3, y denotando con B_{jm} a la m-ésima función base (o B-spline) , es posible representar la función f_j como una combinación lineal de $M_j = t + 3$ B-splines

$$f_j(x_j) = \sum_{m=1}^{M_j} \beta_{jm} B_{jm}(x_j), \tag{15}$$

siendo $\beta_{j1}, \ldots, \beta_{jM_j}$ parámetros desconocidos.

Definiendo el vector de diseño $X_{ji} = (B_{j1}(x_{ji}), \ldots, B_{jM_j}(x_{ji}))$ y el vector de coeficientes $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})$, es posible reescribir el modelo (14) como un modelo lineal

$$y_i = v_i'\gamma + \sum_{j=1}^{l} X_{ji}'\beta_j + \epsilon_i, \ i = 1, \ldots, n. \tag{16}$$

Así representado, el Modelo Aditivo se resume a un modelo lineal, aunque un problema todavía por resolver es el del grado de suavizado a dar a los términos f_j. Este grado de suavizado depende del número de nodos que se especifiquen y del orden de la base de B-splines. En el modelo antes presentado, el orden de la base es igual a 3, y la cantidad de nodos (t_j) no fue especificada.

Un enfoque que se suele seguir en la literatura estadística, y que se sigue en este trabajo, es el de fijar un número suficientemente grande de nodos (entre 20 y 40) y utilizar B-splines cúbicos (como en nuestro caso). Luego, para establecer el grado de suavizado, se define alguna manera de penalizar la curvatura de los términos f_j. Un método pionero en esta línea fue el presentado en Eilers and Marx (1996), el cual se denomina P-splines (penalized splines).

La manera en que los P-splines controlan la curvatura de los términos f_j es utilizando parámetros adicionales (λ_j) para penalizar las diferencias entre coeficientes β_{jm} adyacentes. Luego, la estimación se lleva a cabo mediante Máxima Verosimilitud Penalizada, en donde los términos de penalización vienen dados por

$$P(\lambda_j) = \frac{1}{2}\lambda_j \sum_{m=\delta+1}^{M_j} (\triangle^\delta \beta_{jm})^2,\ \delta = 1, 2, \tag{17}$$

siendo $\triangle^\delta$ el operador diferencia de orden δ.

Partiendo de este enfoque un método similar denominado *Bayesian P-splines* es el que se seguirá en este trabajo. Dicho enfoque da origen a los modelos STAR y es presentado en la siguiente sección.

Definido el MA de forma lineal como en (16), su extensión al Modelo Aditivo Generalizado (GAM) se realiza de la misma forma en la que se extiende el Modelo Lineal Clásico hacia el Modelo Lineal Generalizado (GLM).

Por consiguiente, la única diferencia de especificación entre el GAM y el GLM se encuentra en la definición del predictor o componente sistemática η, que en el caso del GAM es

$$\eta_i = v_i'\gamma + \sum_{j=1}^{l} f_j(x_{ji}) = v_i'\gamma + \sum_{j=1}^{l} X_{ji}'\beta_j,\ i = 1, \ldots, n. \tag{18}$$

Por consiguiente, para modelos multinomiales de respuestas categóricas tanto nominales como ordinales, son válidas en el contexto GAM las definiciones realizadas en la Sección 2.3, con la única salvedad de la componente sistemática.

Entonces, el Modelo Aditivo Logit Multinomial se define como

$$\log\left\{\frac{P(Y_i = r)}{P(Y_i = k)}\right\} = \gamma_{r0} + v_i'\gamma_r + \sum_{j=1}^{l} f_j^{(r)}(x_{ji}) = \gamma_{r0} + v_i'\gamma_r + \sum_{j=1}^{l} X_{ji}'\beta_j^{(r)}, \tag{19}$$

mientras que el Modelo Aditivo Logit Multinomial Acumulativo (para respuesta ordinal) viene dado por

$$\log\left\{\frac{P(Y \leq r)}{P(Y > r)}\right\} = \gamma_{r0} + v_i'\gamma + \sum_{j=1}^{l} f_j(x_{ji}), = \gamma_{r0} + v_i'\gamma + \sum_{j=1}^{l} X_{ji}'\beta_j,\ i = 1, \ldots, n. \tag{20}$$

en donde recordamos que las probabilidades están condicionadas a las covariables.

Los comentarios efectuados en la Sección 2.3 sobre los modelos (11) y (12) son válidos para los modelos (19) y (20). En particular, cabe remarcar que en el caso ordinal los parámetros del vector γ y el coeficiente β_j no dependen de la categoría r, mientras que en el caso nominal sí lo hacen.

Para finalizar la presente sección se exponen dos alternativas para modelar interacciones de forma flexible, entre dos (o mas) variables, dentro de la componente sistemática η. La primer alternativa es utilizar el Modelo de Coeficientes Variables (VCM) introducido por Hastie and Tibshirani (1993). En este marco, los términos $f_j(x_{ji})$ son ampliados a $f_{x_j|z}(x_{ji})z_i$, con el fin de incluir la interacción entre x_{ji} y la variable z_i, donde esta última entra linealmente, es decir, multiplicando al "coeficiente variable" $f_{x_j|z}(x_{ji})$.

La variable x_j suele ser aludida como "modificadora del efecto" de z, ya que el efecto de esta última varía suavemente a lo largo del rango de x_j. En principio z puede ser una variable categórica o continua y provenir del vector de diseño v o del vector de variables continuas x.

Respecto de la representación y estimación del modelo, la inclusión del término de interacción no presenta una complicación, ya que este se puede escribir como

$$f_{x_j|z}(x_{ji})z_i = \sum_{m=1}^{M_{x_j|z}} \beta_{x_j|z}^m z_i B_{jm}(x_{ji}) = z_i X'_{ji}\beta_{x_j|z} = X_i^{\prime x_j|z}\beta_{x_j|z},$$

en donde $\beta_{x_j|z} = (\beta_{x_j|z}^1, \ldots, \beta_{x_j|z}^{M_{x_j|z}})$ y $X_i^{\prime x_j|z} = (z_i B_{j1}(x_{ji}), \ldots, z_i B_{jM_{x_j|z}}(x_{ji}))$ son, respectivamente, el vector de coeficientes y un nuevo vector de diseño (que surge multiplicando el vector X'_{ji} por el escalar z_i).

Lo habitual es que z_j sea categórica, ya que de ser continua existe una forma mas flexible de modelar su interacción con x_{ji}. Por ejemplo, si las variables continuas x_j y x_s interactúan afectando a la respuesta no linealmente, tal interacción puede ser modelada mediante el producto de dos B-splines, es decir

$$f_{j|s}(x_{ji}, x_{si}) = \sum_{m_j=1}^{M_{j|s}} \sum_{m_s=1}^{M_{j|s}} \beta_{jsm_jm_s} B_{jm_1}(x_{ji}) B_{sm_2}(x_{si}).$$

Con esta componente se busca ajustar (no paramétricamente) una superficie a través de un P-spline bidimensional, el cual es aproximado por el producto de dos B-splines unidimensionales (más detalles pueden consultarse en Lang Brezger (2003). Al igual que para el caso unidimensional, esta la componente de interacción se puede escribir en términos de un vector de diseño

$$f_{j|s}(x_{ji}, x_{si}) = X_i^{\prime j|s}\beta_{j|s}.$$

En este caso, $\beta_{j|s} = (\beta_{js11}, \beta_{js11}, \ldots, \beta_{jsM_{j|s}M_{j|s}})$ es un vector de dimensión $(M_{j|s} \times M_{j|s})$ al igual que $X_i^{\prime j|s}$ cuyas componentes son los productos de bases $B_{jm_1}(x_{ji})B_{sm_2}(x_{si})$.

3.2 El Modelo Aditivo Estructurado de Regresión Generalizado (STAR)

Como fuese mencionado en la sección previa, los modelos AM y GAM basados en P-splines pueden ser estimados mediante Máxima Verosimilitud Penalizada, es decir, maximizando la expresión

$$L = \ell(y, \beta_1, \ldots, \beta_l, \gamma) - \lambda_1 \sum_{m=\delta+1}^{M_1} (\triangle^\delta \beta_{1m})^2 - \ldots - \lambda_l \sum_{m=\delta+1}^{M_l} (\triangle^\delta \beta_{lm})^2, \tag{21}$$

con respecto a los parámetros $\beta_1, \ldots, \beta_l$ y γ. La función $\ell(\cdot)$ es la log-verosimilitud, que en el AM se basa en la densidad Normal mientras que en el GLM depende de la distribución elegida para la variable respuesta. Para penalizar el sobre-ajuste, los términos $\lambda_j \sum_{m=\delta+1}^{M_j} (\triangle^\delta \beta_{jm})^2$, $j = 1, \ldots, l$, entran con signo negativo.

En principio se pueden tomar diferencias del orden que se desee, aunque lo habitual es tomar una o dos ($\delta = 1, 2$). Las diferencias de primer orden, $\triangle^1 \beta_{jm} = \beta_{jm} - \beta_{j,m-1}$, castigan los saltos entre coeficientes sucesivos. Por otro lado, las diferencias de orden 2, $\triangle^2 \beta_{jm} = (\beta_{jm} - \beta_{j,m-1}) - (\beta_{j,m-1} - \beta_{j,m-2})$, tienen el fin de penalizar los desvíos de β_{jm} respecto de la tendencia $2\beta_{j,m-1} - \beta_{j,m-2}$.

Desde el punto de vista frecuentista, la estimación puede llevarse a cabo mediante *back-fitting* (Hastie and Tibshirani, 1990) o maximizando directamente de la verosimilitud penalizada (Marx and Eilers, 1998). La selección del parámetro se suavizado λ_j, $j = 1, \ldots, l$, suele basarse en validación cruzada o buscando minimizar el criterio AIC. Estos mecanismos de selección pueden fallar en la práctica al no encontrarse una solución óptima.

En este trabajo utilizamos el enfoque bayesiano, aludido antes como "Bayesian P-splines", en el cual se utilizan las versiones estocásticas de $\triangle^1 \beta_{jm}$ y $\triangle^2 \beta_{jm}$. Estas son especificadas como paseos aleatorios de primer y segundo orden definidos por

$$\beta_{jm} = \beta_{j,m-1} + \varepsilon_{jm} \text{ y } \beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + \varepsilon_{jm}, \tag{22}$$

respectivamente, con $\varepsilon_{jm} \sim N(0, \tau_j^2)$. La varianza τ_j^2 controla el grado de suavizado, menor varianza implica mayor suavizado y viceversa, siendo equivalente a la inversa del parámetro de suavizado λ_j del enfoque tradicional. Cuando estas varianzas son consideradas como constantes (no como variables aleatorias), estamos en el marco del enfoque EB (Empirical Bayes), que es el seguido en este trabajo.

En este contexto se hace necesario definir distribuciones a priori (*priors*) para los parámetros. Para el vector γ se fijan priors difusas independientes, es decir, $p(\gamma) \propto constante$. En el caso del vector β_j de coeficientes β_{jm}, las priors vienen definidas por los procesos estocásticos en (22), las que pueden ser escritas equivalentemente como una *prior global de suavizado*

$$p(\beta_j|\tau_j^2) \propto \exp\left(\frac{1}{2\tau_j^2}\beta_j' K_j \beta j\right), \tag{23}$$

en donde K_j es denominada *matriz de precisión*. Por ejemplo, para un paseo aleatorio de primer orden se tiene

$$\beta_j' K_j \beta_j = \beta_j' \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \beta_j = \sum_{m=2}^{M_j} (\beta_{jm} - \beta_{j,m-1})^2 = \sum_{m=2}^{M_j} (\triangle^1 \beta_{jm})^2. \tag{24}$$

Es importante notar que la matriz de precisión K_j no posee rango completo, dado que $rango(K_j) = M_j - 1$ en el ejemplo anterior y $rango(K_j) = M_j - 2$ para el caso de un paseo aleatorio de segundo orden. Por este motivo la prior (23), que especifica una densidad Normal multivariante para β_j, es una distribución impropia. Este problema hace necesario reparametrizar la componente sistemática (18) y por consiguiente los modelos (19) y (20), para especificarlos como modelos *mixtos*. En este contexto, las componentes del vector γ representan los *efectos fijos* mientras que los elementos de β_j (y por tanto, los términos f_j) son tratados como efectos aleatorios.

Como la reparametrización solo afecta al predictor η, mas específicamente a los coeficientes β_j, las formulaciones que siguen a continuación son válidas tanto para el modelo AM como para el GAM y sus extensiones a modelos multinomiales. Así, tal reparametrización da origen a los Modelos Aditivos Estructurados (STAR), que incluyen al GAM y sus generalizaciones multivariantes. Por motivos de simplicidad, a continuación describiremos el modelo STAR para respuesta multinomial ordinal. Al final de la sección, se extenderán los resultados al caso de respuesta nominal.

Específicamente, la idea consiste en expresar el vector β_j mediante una transformación uno a uno en términos de un vector β_j^{nop} con una prior difusa y otro vector β_j^{pen} con prior Normal i.i.d.. La dimensión de estos dos nuevos vectores depende del rango de la matriz K_j, que llamaremos κ_j. Entonces, si el vector β_j tiene dimensión b_j, su descomposición en una parte no penalizada y otra parte penalizada viene dada por

$$\beta_j = A^{nop}\beta_j^{nop} + A^{pen}\beta_j^{pen}, \tag{25}$$

con una matriz A^{nop} de dimensión $b_j \times (b_j - r\kappa_j)$ y otra matriz A^{pen} de dimensión $b_j \times r\kappa_j$.

Las condiciones requeridas para la descomposición (25) son:

(i) La matriz compuesta $(A_j^{nop}\ A_j^{pen})$ posee rango completo.

(ii) Las matrices A_j^{nop} y A_j^{pen} son ortogonales, es decir, $A_j^{nop}A_j^{pen} = 0$.

(iii) $A_j^{\prime nop}K_jA_j^{nop} = 0$, lo que significa que β_j^{nop} no está penalizado por K_j.

(iv) $A_j^{\prime pen}K_jA_j^{pen} = I$, lo cual implica una prior Normal i.i.d. para β_j^{pen}.

Para más detalles sobre estas condiciones véase Fahrmeir el. al. (2004) y Kneib and Fahrmeir (2004).

Con el cumplimiento de los requisitos (i) a (iv) se obtienen las siguientes distribuciones a priori

$$p(\beta_{jm}^{nop}) \propto constante,\ m = 1, \ldots, b_j - \kappa_j$$

y

$$\beta_j^{pen} = N(0, \tau_j^2 I). \tag{26}$$

La descomposición (25) implica que los términos $f_j(x_{ji})$ también pueden ser descompuestos en una parte penalizada y otra no penalizada:

$$f_j(x_{ji}) = X_{ji}^{\prime}A_j^{nop}\beta_j^{nop} + X_{ji}^{\prime}A_j^{pen}\beta_j^{pen} = \tilde{X}_{ji}^{\prime nop}\beta_j^{nop} + \tilde{X}_{ji}^{\prime nop}\beta_j^{pen}. \tag{27}$$

Por consiguiente, la componente sistemática reparametrizada para el STAR multinomial con respuesta ordinal resulta

$$\eta_i^r = \gamma_{r0} + v_i^{\prime}\gamma + \sum_{j=1}^{l} X_{ji}^{\prime}\beta_j = \gamma_{r0} + v_i^{\prime}\gamma + \sum_{j=1}^{l}(\tilde{X}_{ji}^{\prime nop}\beta_j^{nop} + \tilde{X}_{ji}^{\prime pen}\beta_j^{pen}). \tag{28}$$

Finalmente, para el caso de respuesta nominal se aplica el mismo razonamiento, por lo que la componente sistemática viene dada por

$$\eta_i^r = \gamma_{r0} + v_i^{\prime}\gamma_r + \sum_{j=1}^{l}(\tilde{X}_{ji}^{\prime nop}\beta_j^{nop(r)} + \tilde{X}_{ji}^{\prime pen}\beta_j^{pen(r)}), \tag{29}$$

en donde $\tilde{X}_{ji}^{\prime nop}$ y $\tilde{X}_{ji}^{\prime nop}$ se construyen de la misma forma que para el modelo con respuesta ordinal.

Por último cabe aclarar que los parámetros β_j^{nop} y $\beta_j^{nop(r)}$ son tratados como efectos fijos, mientras que β_j^{pen} y $\beta_j^{pen(r)}$ son efectos aleatorios como lo implica la distribución a priori (26).

3.3. Inferencia con el STAR multinomial

En esta sección se describe el método de estimación del modelo STAR Multinomial de respuesta nominal y STAR Multinomial con respuesta ordinal, cuyas componentes sistemáticas η_i^r fueron definidas en (29) y (28), respectivamente.

Recordemos de la sección 2.3 que, en el caso de respuesta nominal, nuestro objetivo es estimar el modelo que vincula la esperanza (condicional) del vector $y_i = (y_{i1}, \ldots, y_{iq})$, siendo

$$y_{ir} = \begin{cases} 1, \text{ si } Y = r, \ r = 1, \ldots, q = k-1 \\ 0, \text{ en otro caso,} \end{cases}$$

con el vector de predictores $\eta_i = (\eta_i^1, \ldots, \eta_i^q)$ mediante la función respuesta h. Como la esperanza del vector respuesta y_i es el vector de probabilidades $\pi_i = (\pi_{i1}, \ldots, \pi_{iq})$, el modelo queda expresado como

$$\pi_i = (P(y_{i1} = 1), \ldots, P(y_{iq} = 1)) = \left(h(\eta_i^1), \ldots, h(\eta_i^q)\right). \tag{30}$$

Por otra parte, en el caso de respuesta ordinal, lo que se busca es vincular el vector de probabilidades acumuladas (de estar en una determinada categoría o en otra inferior) $\Pi_i = (\Pi_i^{(1)}, \ldots, \Pi_i^{(q)})$ con el vector η_i correspondiente, también mediante la función respuesta h. Así se tiene

$$\Pi_i = (P(Y_i \leq 1), \ldots, P(Y_i \leq q)) = \left(h(\eta_i^1), \ldots, h(\eta_i^q)\right). \tag{31}$$

Entonces, para describir en forma compacta el proceso de estimación de estos dos modelos, el primer paso consiste en definir el vector η_i en términos matriciales. Para ello utilizamos los vectores de diseño

$$\tilde{X}_i^{nop} = (\tilde{X}_{1i}^{\prime nop}, \ldots, \tilde{X}_{li}^{\prime nop}, u_i')' \text{ y } \tilde{X}_i^{pen} = (\tilde{X}_{1i}^{\prime pen}, \ldots, \tilde{X}_{li}^{\prime pen})'$$

los cuales son los mismos tanto para el caso nominal como ordinal.

En segundo lugar escribimos el vector β^{nop} (de parámetros tratados como efectos fijos) y β^{pen} (de efectos aleatorios) de la siguiente manera, para el caso nominal

$$\beta^{nop} = (\beta_1^{\prime nop(1)}, \ldots, \beta_l^{\prime nop(1)}, \gamma_1', \ldots, \beta_1^{\prime nop(q)}, \ldots, \beta_l^{\prime nop(q)}, \gamma_q')'$$

y

$$\beta^{pen} = (\beta_1^{\prime pen(1)}, \ldots, \beta_l^{\prime pen(1)}, \ldots, \beta_1^{\prime pen(q)}, \ldots, \beta_l^{\prime pen(q)})',$$

mientras que para el caso ordinal se tiene

$$\beta^{nop} = (\gamma_{10}, \ldots, \gamma_{l,0}, \beta_1'^{nop}, \ldots, \beta_l'^{nop}, \gamma')'$$

y

$$\beta^{pen} = (\beta_1'^{pen}, \ldots, \beta_l'^{pen})'.$$

En este punto cabe recordar que el vector de efectos aleatorios β^{pen} sigue una distribución normal multivariante según lo establecido en (26), es decir, $\beta^{pen} \sim N(0, \Lambda)$, con

$$\Lambda = \text{bloqdiag}(\Lambda_1^1, \ldots, \Lambda_l^1, \ldots, \Lambda_1^q, \ldots, \Lambda_l^q)$$

para respuesta nominal y con

$$\Lambda = \text{bloqdiag}(\Lambda_1, \ldots, \Lambda_l),$$

para respuesta ordinal. Nótese que Λ contiene los componentes de la varianza de los efectos aleatorios, es decir, contiene los parámetros τ_j^2 que equivalen a la inversa del parámetro de suavizado tradicional.

En tercer lugar, definimos dos matrices que agrupan a los vectores de diseño del siguiente modo

$$Q_i = \begin{pmatrix} \tilde{X}_i^{nop} & & 0 \\ & \ddots & \\ 0 & & \tilde{X}_i^{nop} \end{pmatrix}, \; P_i = \begin{pmatrix} \tilde{X}_i^{pen} & & 0 \\ & \ddots & \\ 0 & & \tilde{X}_i^{pen} \end{pmatrix}, \tag{32}$$

en el caso nominal, mientras que con repuesta ordinal se tiene

$$Q_i = \begin{pmatrix} 1 & & \tilde{X}_i^{nop} \\ & \ddots & \vdots \\ & 1 & \tilde{X}_i^{nop} \end{pmatrix}, \; P_i = \begin{pmatrix} \tilde{X}_i^{pen} \\ \vdots \\ \tilde{X}_i^{pen} \end{pmatrix}. \tag{33}$$

En este punto es posible escribir el vector η_i de dimensión q, tanto para respuesta nominal como para ordinal, de la siguiente forma

$$\eta_i = Q_i \beta^{unp} + P_i \beta^{pen}.$$

Finalmente, se define el modelo en forma matricial para todas las observaciones i de la siguiente manera:

$$\eta = Q\beta^{unp} + P\beta^{pen},$$

en donde η, Q y P están dadas por

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \; Q = \begin{pmatrix} Q_1 \\ \vdots \\ Q_n \end{pmatrix}, \; P = \begin{pmatrix} P_1 \\ \vdots \\ P_n \end{pmatrix}.$$

El mecanismo de estimación se compone de dos etapas que se alternan sucesivamente. Una etapa consiste en estimar los coeficientes en β^{nop} y β^{pen} dados los componentes de la varianza en Λ. La otra etapa es el recíproco de la anterior, es decir, estimar Λ dados los valores de los coeficientes.

La primer etapa computa los estimadores de β^{nop} y β^{pen} maximizando la distribución a posteriori

$$p(\beta^{nop}, \beta^{pen}|y) \propto L(\beta^{nop}, \beta^{pen}) p(\beta^{nop}) p(\beta^{pen}),$$

en donde $L(\beta^{nop}, \beta^{pen})$ es la verosimilitud, cuya forma depende de la función respuesta h y del tipo específico de modelo (con respuesta nominal u ordinal). Tomando logaritmo a la expresión anterior y teniendo en cuenta la prior difusa asignada a β^{nop} se obtiene la expresión

$$\ell_{pen}(\beta^{nop}, \beta^{pen}) = \ell(\beta^{nop}\beta^{pen}) - \frac{1}{2}\beta'^{pen}\Lambda^{-1}\beta'^{pen}, \tag{34}$$

la cual puede ser maximizada respecto de β^{nop} y β^{pen}. Es interesante notar que (34) tiene la forma de una log-verosimilitud penalizada, por lo que el problema de estimación de esta etapa (dado Λ) es equivalente al de Máxima Verosimilitud Penalizada.

La obtención del máximo para (34) puede ser acometida mediante un algoritmo del tipo *Fisher Scoring* . Dicho algoritmo se puede reformular a través de Mínimos Cuadrados Ponderado Iterados (IWLS), resultando en el siguiente sistema de ecuaciones

$$\begin{pmatrix} Q'WQ & Q'WP \\ P'WQ & P'WP + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \beta^{nop} \\ \beta^{pen} \end{pmatrix} = \begin{pmatrix} Q'W\tilde{y} \\ P'W\tilde{y} \end{pmatrix}, \tag{35}$$

el cual debe ser resuelto en cada iteración. La matriz $W = D\Sigma D^{-1}$ posee una estructura diagonal en bloques, constituida por las matrices $D = \text{bloqdiag}(D_1 \; \ldots \; D_n)$, $\Sigma = \text{bloqdiag}(\Sigma_1 \; \ldots \; \Sigma_n)$ y las matrices

$$D_i = \frac{\partial h(\eta_i)}{\partial \eta} = \begin{pmatrix} \frac{\partial h^{(1)}(\eta_i)}{\partial \eta^{(1)}} & \cdots & \frac{\partial h^{(q)}(\eta_i)}{\partial \eta^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h^{(1)}(\eta_i)}{\partial \eta^{(q)}} & \cdots & \frac{\partial h^{(q)}(\eta_i)}{\partial \eta^{(q)}} \end{pmatrix} \tag{36}$$

y

$$\Sigma_i = cov(y_i = \begin{pmatrix} \pi_i^{(1)}(1-\pi_i^{(1)}) & -\pi_i^{(1)}\pi_i^{(2)} & \dots & -\pi_i^{(1)}\pi_i^{(q)} \\ -\pi_i^{(1)}\pi_i^{(2)} & \ddots & & \vdots \\ \vdots & & \ddots & -\pi_i^{(q-1)}\pi_i^{(q)} \\ -\pi_i^{(1)}\pi_i^{(q)} & \dots & -\pi_i^{(q-1)}\pi_i^{(q)} & \pi_i^{(q)}(1-\pi_i^{(q)}) \end{pmatrix}. \tag{37}$$

La respuesta transformada $\tilde{y}$ viene definida por

$$\tilde{y} = \hat{\eta} + (D^{-1})'(y - \pi).$$

Paralelamente, la segunda etapa consistente en la estimación de los componentes de la varianza en Λ. En definitiva el problema es análogo a estimar la varianza de los efectos aleatorios del modelo. Un método tradicional para ello es el de Máxima Verosimilitud Restringida (REML), el cual se basa en por residuos que quedan luego de la estimación de los efectos fijos. Como estos residuos no existen en los modelos generalizados que estamos considerando, entonces se debe utilizar un método equivalente al REML, el cual consiste en maximizar la verosimilitud marginal

$$L^*(\Lambda) = \int L(\beta^{nop}, \beta^{pen}, \Lambda) d\beta^{nop} d\beta^{pen}. \tag{38}$$

En general es necesario realizar una aproximación cuadrática de $L(\beta^{nop}, \beta^{pen}, \Lambda)$ para evaluar la integral en (38). Tal aproximación da origen a la siguiente log-verosimilitud restringida

$$\ell^*(\Lambda) \approx \frac{1}{2}\log(|V|) - \frac{1}{2}\log(|Q'V^{-1}Q|) - \frac{1}{2}(\tilde{y} - Q\beta^{nop})'V^{-1}(\tilde{y} - Q\beta^{nop}), \tag{39}$$

siendo $V = W^{-1} + P'\Lambda P$ una aproximación a la covarianza de $\tilde{y}\beta^{pen}$. La expresión (39) puede ser maximizada mediante Fisher Scoring o Newton Raphson. Para mas detalles del proceso de estimación véase Kneib and Fahrmeir (2004) y Fahrmeir et. al. (2004).

Una vez obtenidos estimadores de β^{nop} y β^{pen} es posible escribir el estimador de la función f_j, aplicando la expresión (27), como

$$\hat{f}_j(x_{ji}) = \tilde{X}_{ji}^{\prime nop}\hat{\beta}_j^{nop} + \tilde{X}_{ji}^{\prime pen}\hat{\beta}_j^{pen}.$$

También es posible construir intervalos de confianza para f_j, utilizando la matriz de covarianzas de los coeficientes estimados $\hat{\beta}^{nop}$ y $\hat{\beta}^{pen}$. Esta matriz de covarianzas viene dada por H^{-1} y siendo H la matriz de coeficientes ubicada en el lado izquierdo de (35). Con estos elementos se obtiene el error estándar para $\hat{f}_j$:

$$se(\hat{f}_j(x_{ji})) = \sqrt{(\tilde{X}_{ji}^{\prime nop} \ \tilde{X}_{ji}^{\prime pen})\mathrm{Cov}\left((\hat{\beta}_j^{\ nop})' \ (\hat{\beta}_j^{\ pen})'\right)(\tilde{X}_{ji}^{\prime nop} \ \tilde{X}_{ji}^{\prime pen})'}.$$

Las matrices de covarianzas $\text{Cov}\left((\hat{\beta}_j^{nop})' \ (\hat{\beta}_j^{pen})'\right)$se obtienen de los bloques correspondientes en H^{-1}.

4. Explorando la Prevalencia de Pre-diabetes y Diabetes

En este capítulo se presentan aplicaciones de los modelos de regresión STAR de respuesta multinomial. Lo que se busca es comparar y evaluar modelos de regresión alternativos. Para estimar dichos modelos, tanto en este capítulo como en el siguiente, se empleó el paquete estadístico BayesX (Brezger et. al., 2005).

Específicamente, el presente capítulo deriva algunas conclusiones en base a un análisis exploratorio del fenómeno bajo estudio. El objetivo principal es describir y evaluar factores de riesgo que serán incluidos como covariables en los modelos de regresión, los cuales son analizados con más detalle en el Capítulo 5.

4.1 Origen y descripción de los datos utilizados

La base de datos utilizada en este trabajo, fue realizada en base a una muestra aleatoria representativa de la población adulta (mayores de 18 años) de Galicia, entre los meses de Marzo y Julio del 2004. Para la construcción de la misma fue empleado un procedimiento de muestreo cluster en dos etapas, sobre un directorio de datos perteneciente al Servicio Gallego de Salud (SERGAS) que cubre al 95 % de la población de interés. En primer lugar se seleccionaron al azar Centros de Atención Primaria (CAP) para cada una de las cuatro provincias gallegas (cuyas poblaciones se consideraron independientes). De la población dependiente de cada CAP obtenido, se seleccionaron al azar individuos mayores de 18 años. Los CAP fueron estratificados por tipo de ayuntamiento (rural o urbano; costero o interior) mientras que los individuos lo fueron por edad y sexo. Las mujeres embarazadas no se incluyeron en el estudio.

Las personas fueron contactadas por mail para acordar su participación, siendo sustituidas las no respondientes por un sustituto seleccionado aleatoriamente. La información fue extraída mediante un cuestionario en una entrevista personal realizada en los CPA, seguida de un examen físico y un análisis de sangre.

Las medidas antropométricas fueron tomadas por personal entrenado utilizando técnicas y equipamiento estándar. En el presente trabajo se utilizan dos variables antropométricas: el índice de masa corporal (IMC) y la obesidad central (OC). El IMC se computa como el ratio entre peso (en kilogramos) y la altura (en metros) al cuadrado, mientras que la OC se determina con la circunferencia de la cintura (en centímetros).

El cuestionario también incluye información sobre características socio-demográficas. De las mismas, en este estudio se consideran el sexo, la edad (medida en años) y el nivel educativo. El nivel educativo se clasifica en cinco grupos: iletrados o personas sin educación formal aunque sepan leer y escribir (grupo 1), educación primaria (grupo 2), educación completada

a los 13 o 14 años (grupo 3), educación completada entre los 16 y 19 años o estudios superiores no universitarios (grupo 4) y estudios universitarios (grupo 5). Esta característica se representa por una variable discreta que toma valores desde el 1 (grupo 1 de educación) al 5 (grupo 5).

La variable de interés principal en el estudio es la que clasifica a las personas como diabéticas, pre-diabéticas o normales. Para los fines del presente estudio, las categorías de diabetes fueron construidas de la siguiente manera:

- Diabetes Mellitus (DM): si el anásis de sangre detecta *glucosa basal*$\geq$ 126 mg/ml, o *glucosa 2-h* $\geq$ 200 mg/ml o el individuo sigue un tratamiento para la diabetes.

- Pre-diabetes (PreDM): si el análisis demuestra 110 mg/ml $\leq$ *glucosa basal* $<$ 126 mg/ml y/o 140 mg/ml $\leq$ *glucosa 2-h* $<$ 200 mg/ml.

- Normal: si no se presenta ninguna de las dos categorías previas.

Así definida, tal variable es categórica ordinal con tres categorías, en la cual el estado "normal" es la categoría de menor gravedad y el estado DM es el de mayor gravedad. De ahora en adelante llamaremos *gluco* a dicha variable, la cual toma los valores 1, 2 y 3 para las categorías "normal", Pre-DM y DM, respectivamente.

Como el objetivo del análisis es relacionar la variable *gluco* con las restantes variables antropométricas (*IMC* u *OC*) y socio-demográficas (*sexo*, *edad* y *educación*), en el Cuadro 1 se presenta la media y el desvío estándar de estas variables separándolas por sexo y categorías de *gluco*.

	Normal			**Pre-diabetes (PreDM)**			**Diabetes (DM)**		
Variables	media	desvío	frec.	media	desvío	frec.	media	desvío	frec.
Hombres			855			351			130
IMC	26.678	4.342		28.208	3.604		28.464	3.828	
OC	90.520	11.739		95.264	10.573		96.984	10.498	
edad	36.324	12.148		48.867	15.164		55.416	14.264	
educación	3.485	1.050		2.854	1.123		2.469	1.155	
Mujeres			1197			254			100
IMC	25.693	5.354		28.722	5.346		27.753	4.949	
OC	81.111	13.263		88.448	13.441		89.18	13.891	
edad	38.601	13.874		50.793	15.596		57.231	16.478	
educación	3.499	1.151		2.566	1.146		2.26	1.078	

Cuadro 1: *Valores medios y desvíos estándar de las variables, por niveles de glucosa.*

La separación a priori entre hombres y mujeres, en el contexto del estudio de dibetes, es una recomendación bastante común en la literatura médica, dada la significativa diferencia que existe entre estos dos grupos.

Típicamente la edad tiene un gran peso sobre la probabilidad de presentar problemas de elevada glucosa, por ello siempre es tenida en cuenta en los estudios. En la muestra estudiada, esta relación positiva entre edad y niveles de glucosa se refleja, tanto en hombres como en mujeres, en el aumento de la media de edad al pasar de categorías de menor glucosa a otras con mayor glucosa. Algo similar sucede con la educación, aunque en sentido inverso, a menor nivel educativo parece haber más posibilidades de presentar alta glucosa.

A diferencia de la edad y el nivel de educación, las cuales no son características susceptibles de ser influidas por las campañas preventivas y la recomendación médica, el IMC y la OC sí son variables que las personas pueden afectar mediante su conducta alimenticia. Estas dos variables, que tienen como finalidad medir el sobrepeso o nivel de obesidad de los individuos, representan un factor importante en el riesgo de padecer niveles altos de glucosa. En cuanto a sus valores medios, el IMC y la OC presentan claras diferencias solo entre la categoría normal y las otras dos (DM y PreDM), tanto en hombres como en mujeres, aunque para estas últimas tales diferencias son mas marcadas.

4.2 Prevalencia de diabetes y pre-diabetes según sexo, edad, IMC y OC

Como punto de partida en el análisis de los datos, en esta sección se relaciona a la prevalencia de DM y PreDM con las variables edad, IMC y OC, separando entre hombres y mujeres. Para ello se utilizan modelos de regresión simple que relacionan a la variable categórica *gluco* con cada una de las variables explicativas. Lo que se busca con esto es, en principio, ilustrar la gran importancia que tiene la edad como determinante del nivel de glucosa, en segundo lugar, describir el efecto del IMC y la OC sobre la prevalencia cuando son tomadas aisladamente y, en tercer lugar, hacer notar las diferencias existentes entre hombres y mujeres, respecto de las mencionadas relaciones.

Los modelos específicos a utilizar son STAR y GLM multinomiales para respuesta *nominal*, descriptos en las secciones 3.2 y 2.3 respectivamente, utilizando una función link logística (dando origen a los modelos Logit Multinomiales). La utilización de estas dos clases de modelos, de las que el GLM es una sub-clase del STAR, brinda la posibilidad de compararlas, detectando las diferencias entre ambas y haciendo notar cuándo se justifica el uso de cada una.

Si bien la variable *gluco* posee una naturaleza ordinal, se comienza tratándola como nominal con la finalidad de estimar las prevalencias de pre-diabetes y diabetes en relación a

la categoría de referencia, que en este caso viene dada por el estado "normal".

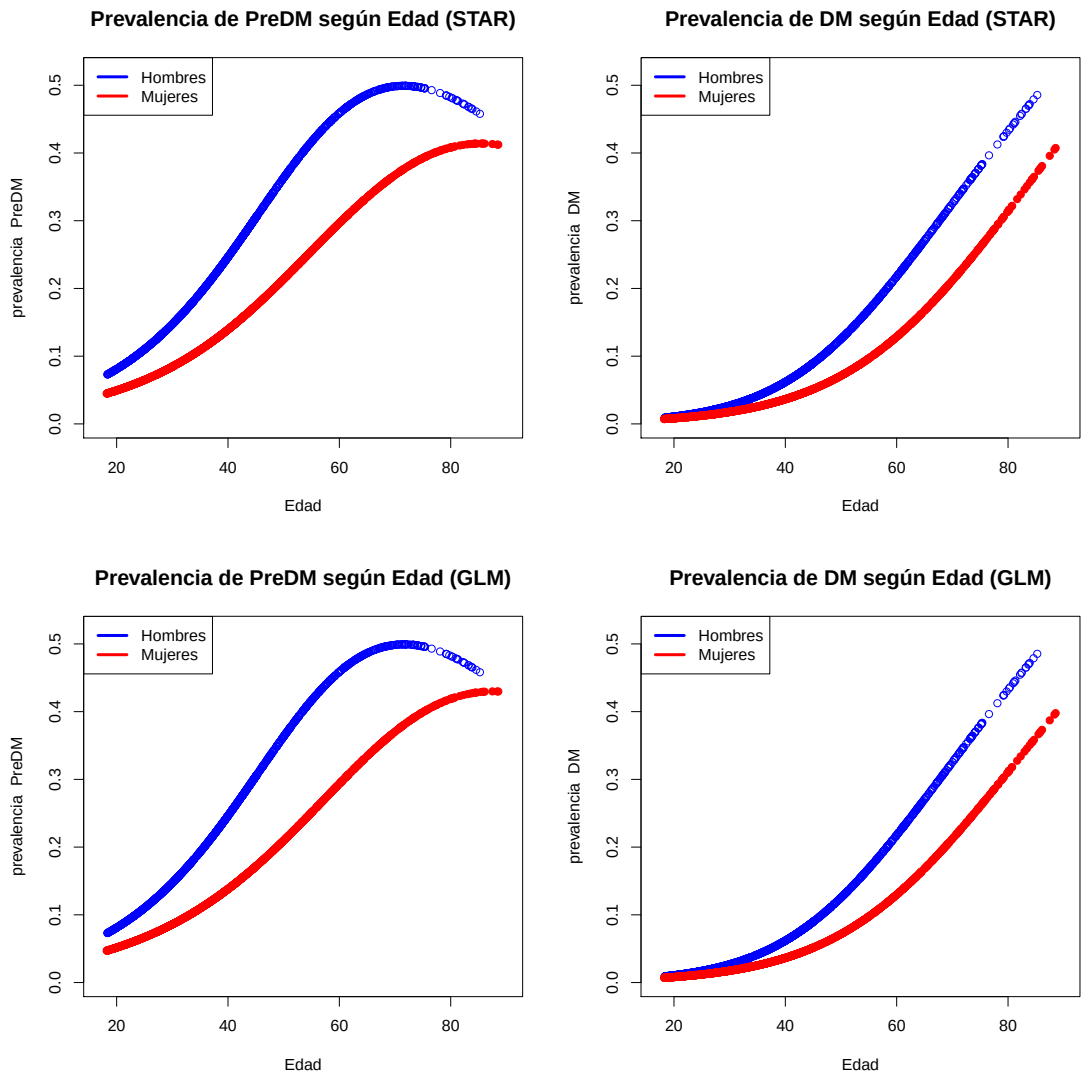

Figura 1: *Relaciónes entre la edad y las prevalencias de pre-diabetes (izquierda) y de diabetes (derecha), ajustadas con modelos STAR (arriba) y GLM (abajo). Las prevalencias para hombres se grefican con puntos huecos azules, mientras que para las mujeres con puntos sólidos rojos.*

Para ajustar los modelos, las muestras fueron separadas en hombres y mujeres, estimándose por lo tanto seis modelos STAR y otros seis GLM. Como se mencionó anteriormente, con este tipo de modelos se estiman dos prevalencias, una se define como la probabilidad de pasar de la categoría de referencia ("normal") a la de pre-diabetes (PreDM) y la otra

representa la probabilidad de saltar de la categoría de referencia a la de diabetes (DM).

La Figura 1 presenta las prevalencias de pre-diabetes y diabetes estimadas para cada valor de la variable edad. En primer lugar es importante notar la gran incidencia del paso de la edad sobre las prevalencias. Aunque esta incidencia de la edad se explica en parte por otras características relacionadas con la misma y no incluidas en este modelo (como ser el IMC u OC y el nivel educativo), el solo paso del tiempo representa en sí mismo un factor de peso.

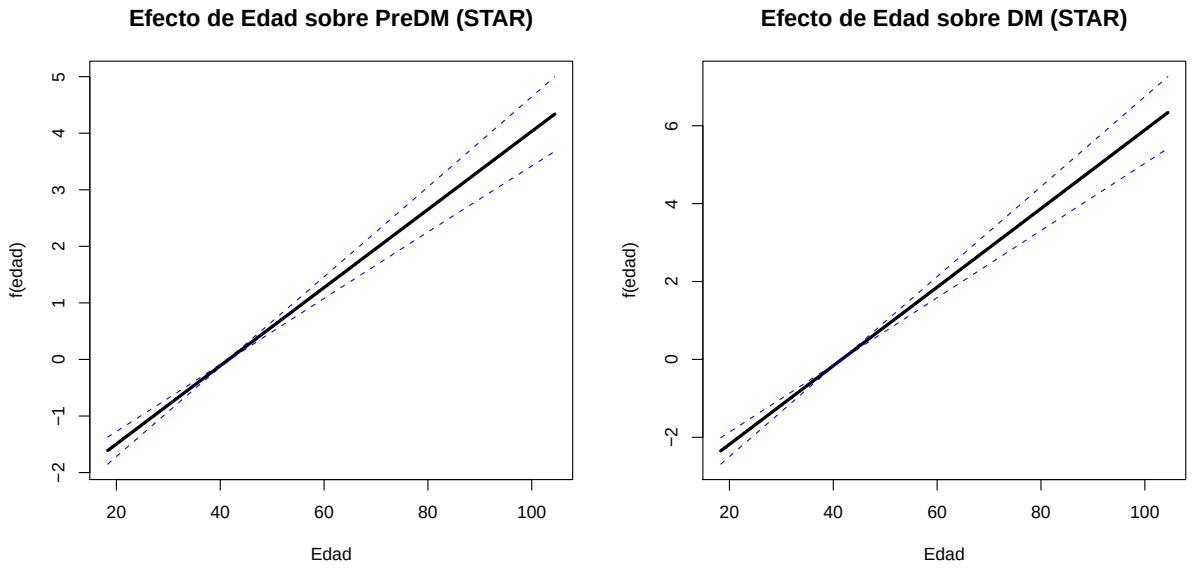

Figura 2: *Efectos centrados de la edad sobre la prevalencia, estimados con el modelo STAR para los hombres. Como se aprecia, los efectos de la edad son lineales.*

Otra cuestión importante viene dada por las marcadas diferencias existentes entre hombres y mujeres. Tanto en caso de diabetes como de pre-diabetes, los hombres presentan un mayor nivel de prevalencia que las mujeres, para todas las edades. Además, las prevalencias máximas en los hombres alcanzan el 50 %, mientras que para las mujeres rondan el 40 %. Respecto de la prevalencia de pre-diabetes, para los hombres se registra un punto de inflexión entre los 65 y 70 años de edad, mientras que para las mujeres tal prevalencia alcanza su máximo alrededor de los 85 años.

En cuanto a la comparación entre el modelo STAR y el GLM, las estimaciones resultan prácticamente las mismas. Esto se debe a que el efecto de la variable edad sobre las prevalencias es lineal. A modo ilustrativo, la Figura 2 muestra la estimación STAR de dicho efecto para las dos prevalencias en el caso de los hombres, pudiéndose observar la linealidad de los

mismos. Además, en los gráficos se pueden ver los intervalos de confianza del 95 % para los efectos, es decir, los cuantiles a posteriori del 2.5 % y 97.5 %.

De forma análoga a lo realizado para la variable edad, la Figura 3 ayuda a explorar la relción entre el IMC y la prevalencia de DM y Pre-DM, distinguiendo entre hombres y mujeres.

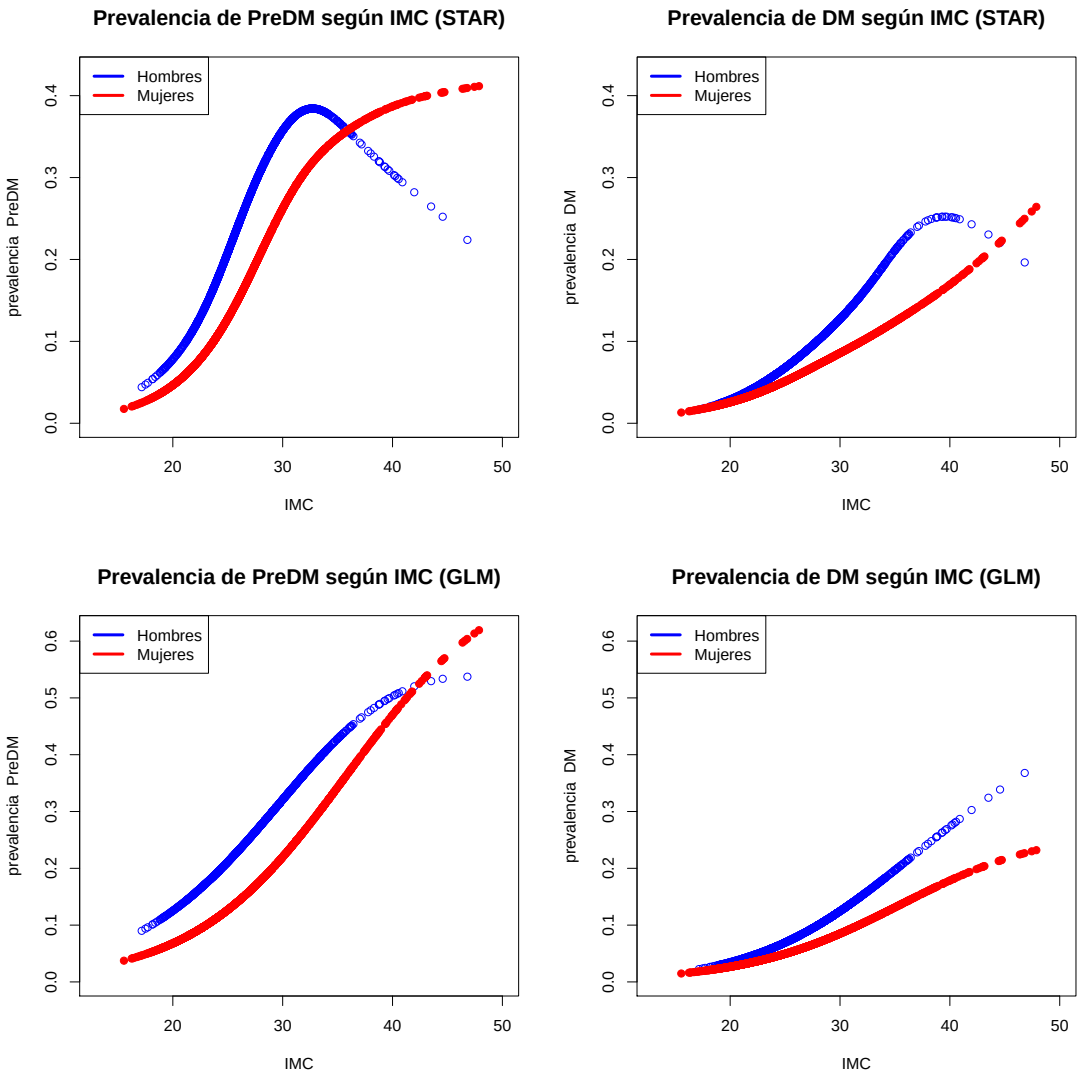

Figura 3: *Relaciónes entre el IMC y las prevalencias de pre-diabetes (izquierda) y de diabetes (derecha), ajustadas con modelos STAR (arriba) y GLM (abajo).*

Lo primero que conviene notar es la existencia de marcadas diferencias entre los resultados

obtenidos con el modelo STAR respecto de los arrojados por el GLM. Estas son debidas a que los efectos del IMC sobre las dos prevalencias no son lineales, como lo eran en el caso de la edad, lo cual se aprecia an la Figura 4. Por este motivo, los comentarios se harán sobre las estimaciones basadas en los STAR.

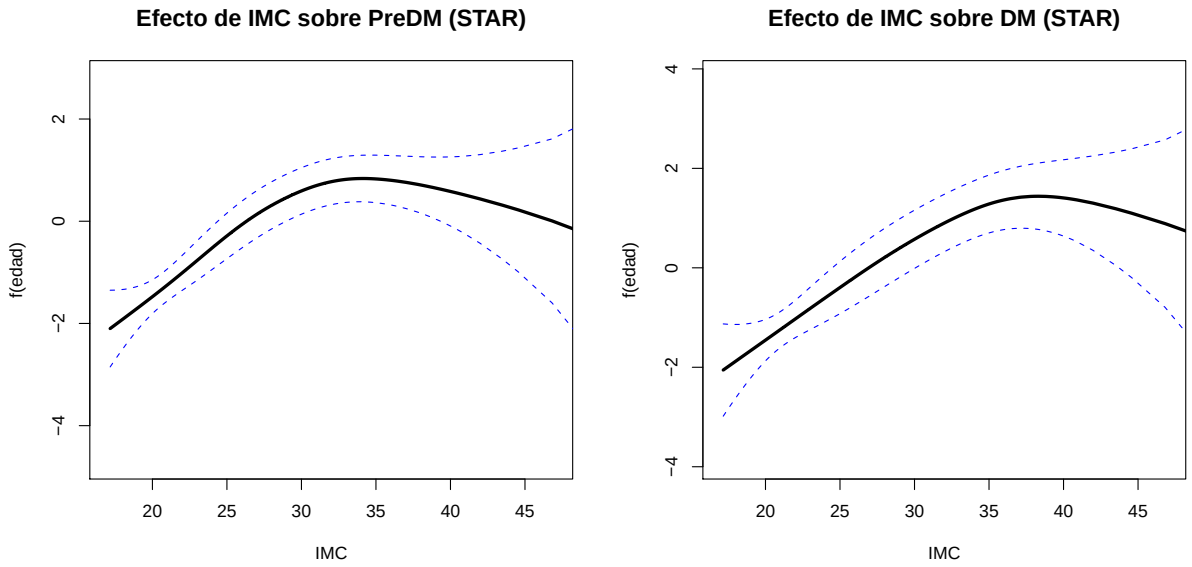

Figura 4: *Efectos centrados del IMC sobre la prevalencia, estimados con el modelo STAR para los hombres. En este caso los efectos son no lineales.*

Al igual que con la edad, se vuelven a apreciar brechas entre las prevalencias de los hombres y la de las mujeres, las que difieren tanto en nivel como en su evolución a medida que aumenta el IMC. Para ambas prevalencias los hombres presentan un punto de inflexión (en IMC = 32 para Pre-DM y en IMC = 39 para DM, aproximadamente), mientras que las mujeres no lo tienen. Además, en el caso de pre-diabetes, los hombres superan en prevalencia a las mujeres en los valores iniciales del IMC hasta que éste llega a 35, a partir de allí la prevalencia en mujeres es superior a la de hombres.

Finalmemte, en los que respecta la incidencia de la OC, la Figura 5 expone la relación entre esta variable y la prevalencia de pre-diabetes y de diabetes, para hombres y mujeres. Nuevamente los modelos STAR presentan estimaciones que difieren con las obtenidas con los GLM (aunque principalmente para la prevalencia de PreDM).

En esta ocasión parece haber cierta similitud en la prevalencia de DM entre hombres y mujeres, solo para valores muy altos de OC parece dispararse la probabilidad para los hombres.

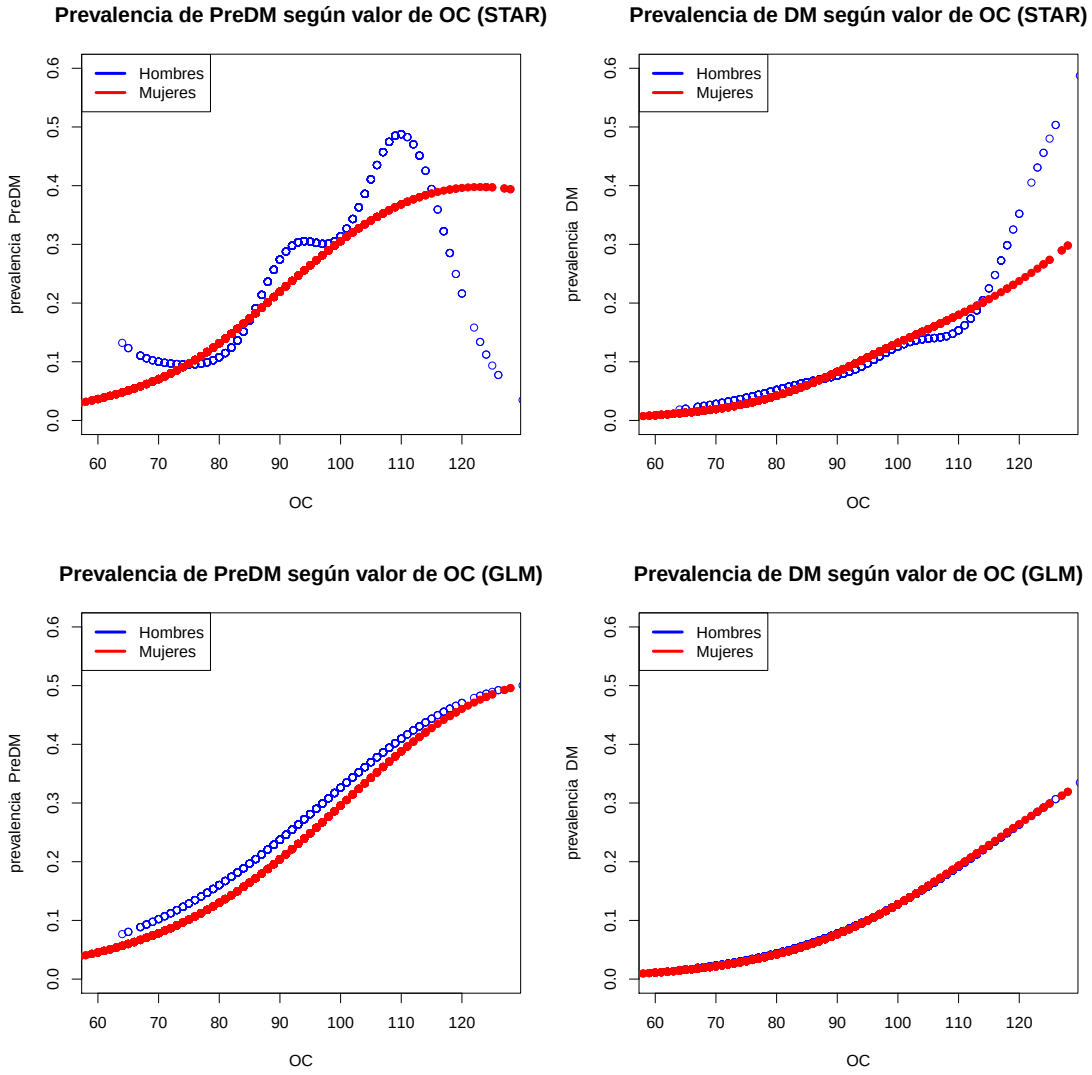

Figura 5: *Relaciónes entre la OC y las prevalencias de pre-diabetes (izquierda) y de diabetes (derecha), ajustadas con modelos STAR (arriba) y GLM (abajo).*

Respecto de la prevalencia de PreDM, mientras que para las mujeres presenta un comportamiento suave (con un máximo de 40 % para una OC de 120), para los hombres muestra una evolución irregular, registrándose una meseta inicial (entre valores de OC de 65 y 80) y otra meseta intermedia (entre los 90 y 100 de OC).

Se concluye de esta sección, que es conveniente separar a hombres y mujeres para llevar a cabo los análisis. Además es recomendable el uso de modelos STAR, para tener en cuenta posibles efectos no lineales de los factores que afectan las prevalencias.

4.3 Evaluando factores de riesgo para la DM y PreDM

En esta sección se expanden los modelos STAR multinomiales, ajustados en la sección previa, a través de la inclusión simultánea de varias covariables. El objetivo general es delinear modelos, posiblemente alternativos, con la capacidad de predecir los estados de diabetes y pre-diabetes y, por tanto, ser útiles en la clasificación de nuevos casos. En particular, lo que se busca es evaluar el IMC y la OC como factores de riesgo, indagando sobre si pueden ser sustitutos o complementarios a la hora de explicar las prevalencias.

Por ahora, seguimos tratando a la variable *gluco* como nominal. Esto facilita analizar por separado a las prevalencias de Pre-DM y DM. Más adelante, en el capítulo siguiente, se abordarán los modelos que explican a *gluco* teniendo en consideración su carácter ordinal.

El enfoque a seguir en esta sección es la comparación de diversos modelos mediante los criterios AIC (*Akaike Information Criterion*), BIC (*Bayesian Information Criterion*) y GCV (*Generalized Cross-Validation*). Si bien este tipo de comparaciones es de carácter descriptivo, ya que no conforman un procedimiento de contraste de hipótesis, se recae provisoriamente en ellos dada la ausencia de procedimientos de contraste para este tipo de modelos. De cualquier manera, mas adelante recurrimos al análisis de curvas ROC para la comparación final de los modelos.

En el Cuadro 2 se presentan los diversos modelos ajustados, los cuales quedan definidos por el predictor η (por simplicidad se omite el intercepto). Los valores de AIC, BIC y GCV se exponen para los dos modelos básicos, es decir, para los que solo incluyen la edad (modelos 1 y 8, con $\eta = f(edad)$) y para los que incluyen edad y educación (modelos 4 y 11, con $\eta = f(edad) + \beta educ$). Para los demás modelos se reporta la variación del AIC, BIC y GCV ($\triangle$AIC, $\triangle$BIC y $\triangle$GCV) computada respecto del modelo básico que corresponda. También son incluidos los grados de libertad de cada modelo estimado.

La primer variable a considerar es la educación, la cual parece tener un aceptable poder explicativo si se tiene en cuenta los criterios AIC, BIC y GCV (excepto el BIC para el caso de los hombres), junto con el hecho de que consta con solo 5 niveles y un efecto lineal constante para cada uno de ellos.

Centrándonos en el grupo de las mujeres, tanto el IMC como la OC aportan un sustancial poder explicativo si se incluyen separadamente, como lo sugieren las variaciones negativas de los tres criterios en los modelos 2, 3, 5 y 6. Es interesante notar que, para las mujeres, la educación está relacionada con el IMC y la OC. Tal relación se aprecia al comparar los modelos 2 y 5 (para el IMC) y el 3 con el 6 (para la OC). En ambos casos, el hecho de controlar por la educación hace disminuir el poder explicativo de estos indicadores de sobrepeso. La explicación que se encuentra es que una mayor educación produce una mayor conciencia en

Modelo	**Predictor** ($\eta = \gamma_{r0} + \ldots$)	**g.l.**	**AIC** ($\triangle$AIC)	**BIC** ($\triangle$BIC)	**GCV** ($\triangle$GCV)
	Mujeres				
1	$f(edad)$	4.3423	**1840.02**	**1863.24**	**1.1742**
2	$f(edad) + f(IMC)$	8.0848	(-65.06)	(-45.06)	(-0.052)
3	$f(edad) + f(OC)$	6.4439	(-59.75)	(-48.52)	(-0.0441)
4	$f(edad) + \gamma_{r1} educ$	6.4646	**1824.09**	**1858.65**	**1.1580**
5	$f(edad) + \gamma_{r1} educ + f(IMC)$	9.6967	(-54.7)	(-37.41)	(-0.0438)
6	$f(edad) + \gamma_{r1} educ + f(OC)$	8.3876	(-49.16)	(-38.18)	(-0.036)
7	$f(edad) + \gamma_{r1} educ + f(IMC) + f(OC)$	11.7509	(-60.75)	(-32.48)	(-0.0532)
	Hombres				
8	$f(edad)$	4.0127	**1980.21**	**2001.06**	**1.4673**
9	$f(edad) + f(IMC)$	7.4561	(-30.41)	(-12.51)	(-0.0351)
10	$f(edad) + f(OC)$	11.3579	(-35.43)	(2.75)	(-0.053)
11	$f(edad) + \gamma_{r1} educ$	6.1318	**1976.15**	**2008.02**	**1.4565**
12	$f(edad) + \gamma_{r1} educ + f(IMC)$	9.4804	(-30.2)	(-12.8)	(-0.0346)
13	$f(edad) + \gamma_{r1} educ + f(OC)$	13.2724	(-34.96)	(2.15)	(-0.0517)
14	$f(edad) + \gamma_{r1} educ + f(IMC) + f(OC)$	15.1296	(-38.84)	(7.93)	(-0.0612)

Cuadro 2: *Comparativa de modelos de complejidad creciente, para hombres y mujeres.*

el cuidado de la alimentación, haciendo menos probable sufrir de sobrepeso.

Por el lado de los hombres, el IMC y la OC continúan teniendo peso al explicar las prevalencias, aunque en menor medida que para las mujeres. Un punto a notar es la mayor cantidad relativa de grados de libertad que requiere la inclusión de la variable OC. Esta es la razón por la cual el criterio BIC no disminuye al incluir la OC en los modelos 10 y 13. En contraste con el caso de las mujeres, la educación no afecta el poder explicativo del IMC y la OC (compárese modelos 9 y 12 por un lado, y modelos 10 y 13 por otro). Ello implica que para los hombres hay una baja relación entre la educación y estas dos variables.

Por último queda por analizar la cuestión de si el IMC y la OC son covariables definitivamente sustitutas o se complementan en alguna medida. A priori, ambas variables son utilizadas como indicadores del sobrepeso que presenta una persona y, por lo tanto, son concebidas como sustitutas. Sin embargo, como se aprecia en el Cuadro 2, la utilización de una u otra no resulta exactamente lo mismo.

Un primer paso para analizar la cuestión es comparar el modelo 7 con el 5 y el 6 en el grupo de mujeres, y el modelo 14 con el 12 y 13 en el caso de los hombres. Según los

criterios AIC y BIC, el uso simultáneo de IMC y OC aporta algo mas de poder explicativo respecto de su inclusión por separado. Pero este mayor poder explicativo es relativamente escaso si se lo compara con el aporte individual que tienen el IMC y la OC cuando son usadas separadamente. Por ejemplo, al pasar del modelo 4 al 5 (incluyendo solo el IMC) hace reducir el criterio AIC en 54.7 puntos, mientras que el paso del modelo 4 al 7 (incluyendo tanto IMC como OC) reduce el AIC en 60.7 puntos. Ello significa que con la utilización conjunta de IMC y OC solo se gana una reducción de 6 puntos (del criterio AIC) respecto de usar solo el IMC.

Los comentarios previos dejan dudas en relación al uso conjunto de IMC y OC como covariables. En el capítulo siguiente se vuelve a indagar sobre este tema, al analizar los efectos estimados de cada variable.

5. Modelos con respuesta nominal y ordinal en el análisis de la prevalencia

En este capítulo se presentan los efectos estimados, sobre la prevalencia de DM y Pre-DM, que ejercen los factores de riesgo IMC y OC. En primer lugar, la Sección 5.1 trata con los modelos STAR multinomiales con respuesta nominal. En segundo lugar, en la Sección 5.2 se explota la naturaleza ordinal de la variable *gluco* empleando los mismos modelos STAR pero con respuesta ordinal. Posteriormente, en la Sección 5.3, se construyen curvas ROC para completar el análisis y evaluación de la bondad de los modelos estimados.

5.1 IMC y OC como factores de riesgo para la diabetes y pre-diabetes

Esta sección profundiza en los efectos estimados del IMC y la OC para los modelos 5, 6, 7, 12, 13 y 14 definidos en el Cuadro 2 de la Sección 4.3. Como fue mencionado antes, dichos modelos pertenecen a la clase STAR, siendo especificados como modelos Logit Multinomiales con respuesta nominal. La finalidad que se persigue ahora no es la justificación de algún modelo solo en términos de su bondad de ajuste. Lo que se busca es ilustrar cómo es posible evaluar la idoneidad de un modelo en base a los efectos estimados de los factores de interés.

Para los modelos 5 y 12, que incluyen la educación, la edad y el IMC se tiene (compárese con la expresión (19) en la Sección 3.1):

$$\log\left\{\frac{P(Y_i=r)}{P(Y_i=k)}\right\} = \gamma_{r0} + \gamma_{r1}educ_i + f^{(r)}_{edad}(edad_i) + f^{(r)}_{IMC}(IMC_i) = \eta^r_i, \tag{40}$$

con $i = 1, \ldots, n$, en donde $r \in \{\text{PreDM, DM}\}$, la categoría de referencia es $k =$ "normal" y la probabilidad de pertenencia a la categoría r es

$$P(Y_i = r) = \frac{\exp(\eta^r_i)}{1 + \exp(\eta^{PreDM}_i) + \exp(\eta^{DM}_i)}. \tag{41}$$

Al término del lado izquierdo en (40) lo llamaremos *log-odds*. Por lo tanto, al aumentar el IMC varía el término $f^{(r)}_{IMC}(IMC_i)$ afectando directamente al log-odds.

En la Figura 6 se puede apreciar el efecto del IMC en el grupo de hombres (modelo 12), tanto sobre el log-odds en donde $r =$ PreDM como sobre el log-odds en el que $r = DM$. Recuérdese que la categoría de referencia siempre es $k =$ "normal". El efecto sobre el log-odds de PreDM es no lineal, creciente hasta un IMC de 33 y estable de allí en más. En cuanto a la DM, el efecto del IMC es lineal y creciente.

Para las mujeres, los efectos del IMC (correspondientes al modelo 5) se visualizan en la Figura 7. En este caso, los dos efectos son crecientes y levemente no lineales. Para la Pre-DM

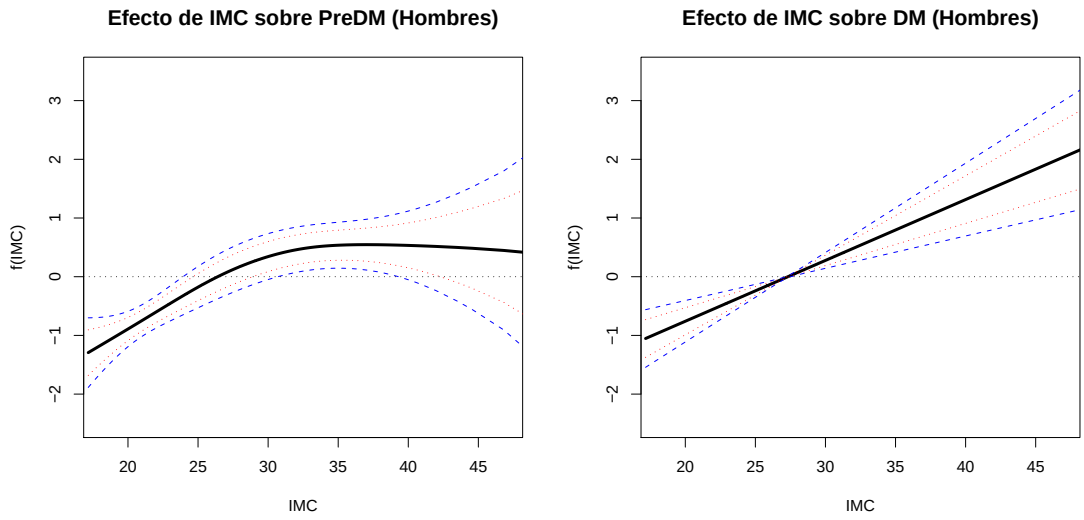

Figura 6: *Efectos centrados del IMC sobre el log-odds de PreDM y DM, estimados para los hombres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

el efecto tiene una tasa de crecimiento positiva pero suavemente decreciente, mientras que para la DM se observa una tasa positiva y levemente creciente. Nuevamente queda reflejado

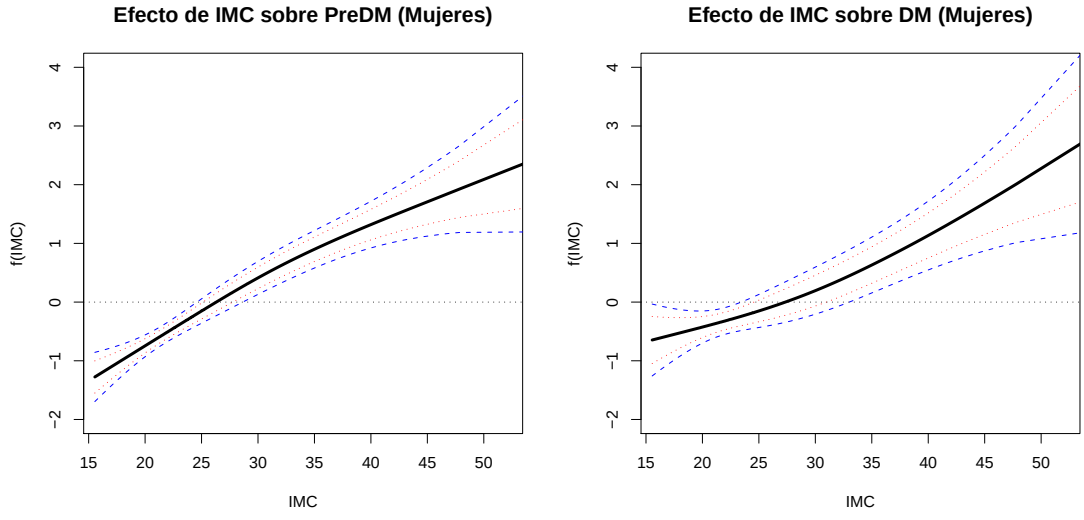

Figura 7: *Efectos centrados del IMC sobre el log-odds de PreDM y DM, estimados para las mujeres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

cómo el fenómeno estudiado difiere entre hombres y mujeres, aunque en esta oportunidad la diferencia sustancial se observa para el log-odds de la Pre-DM.

La expresión del log-odds para los modelos 6 y 13, que incluyen la OC en vez del IMC, viene dada por

$$\log\left\{\frac{P(Y_i = r)}{P(Y_i = k)}\right\} = \gamma_{r0} + \gamma_{r1} educ_i + f^{(r)}_{edad}(edad_i) + f^{(r)}_{OC}(OC_i),\ i = 1, \ldots, n, \qquad (42)$$

en donde, al igual que antes, $r \in \{\text{PreDM, DM}\}$ y la categoría de referencia es $k =$ "normal".

La Figura 8 expone las gráficas del efecto de la OC sobre las log-odds de PreDM y DM, en el caso de los hombres (modelo 13). Según se observa en la gráfica de la izquierda, el

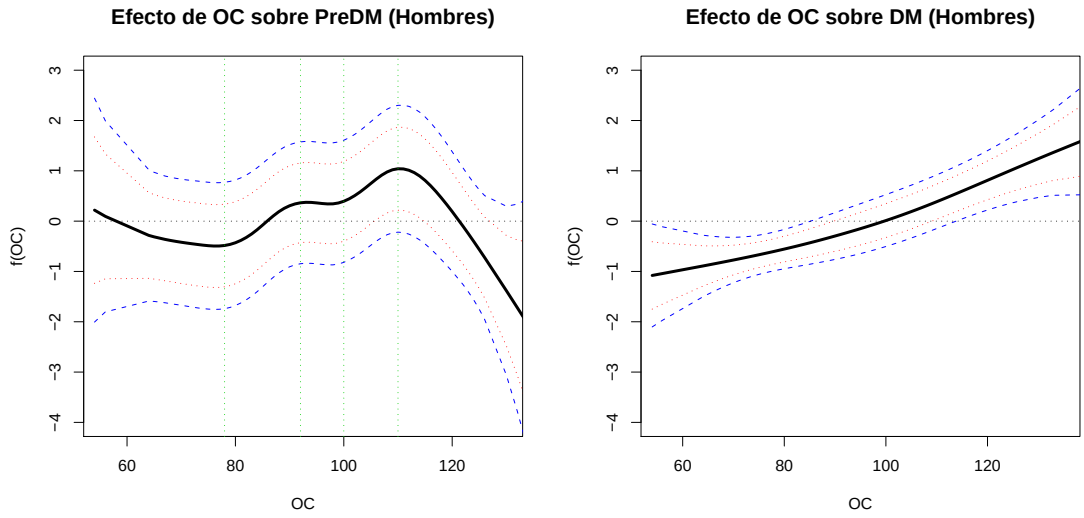

Figura 8: *Efectos centrados de la OC sobre el log-odds de PreDM y DM, estimados para los hombres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos), además de rectas verticales indicando puntos de inflección.*

efecto de la OC, sobre el log-odds de PreDM, muestra una no linealidad mas compleja de lo que se venía observando hasta ahora. Entre los 55 y 78 centímetros se aprecia un efecto algo decreciente y predominantemente negativo. Desde los 78 a los 92 centímetros, el efecto se vuelve creciente y pasa de negativo a positivo a los 86 cm. aproximadamente. Entre los 92 y los 100 cm. sigue siendo positivo pero relativamente estable. Al superar los 100 cm. el efecto vuelve a experimentar un crecimiento hasta llegar a los 110 cm., en donde comienza a decrecer volviéndose negativo al traspasarse los 121 cm. de OC. Por otro lado, el efecto de la OC sobre DM es creciente a tasa prácticamente constante.

Concentrándonos ahora en el efecto de la OC sobre los log-odds para el grupo de mujeres, los mismos se presentan en la Figura 9. En esta ocasión, ambos efectos son lineales

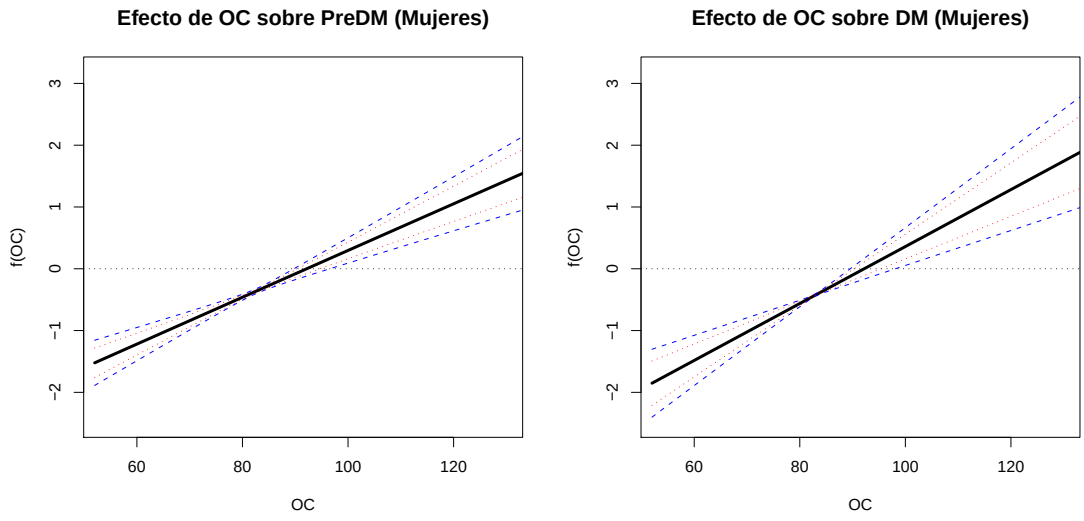

Figura 9: *Efectos centrados de la OC sobre el log-odds de PreDM y DM, estimados para las mujeres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

y crecientes, con una tasa de crecimiento similar (levemente superior para DM). Resulta interesante mencionar que los efectos estimados de la variable edad también son lineales en este modelo. Por ello, es recomendable en este caso utilizar el modelo GLM si se desea ganar precisión (es decir, eficiencia) en las estimaciones de los efectos. Esta es una buena oportunidad para justificar la ventaja de utilizar un modelo flexible como el GAMM, en el sentido de que es capaz ajustar muy bien un modelo lineal relativamente complejo.

Para ilustrar mejor lo mencionado, se pueden comparar las estimaciones GAMM contra las GLM mediante los criterios AIC, BIC y GCV. Para el GAMM, los criterios AIC, BIC y GCV son de 1774.93, 1819.77 y 1.1213, por su parte, para el GLM los respectivos valores son 1774.98, 1817.75 y 1.1224. Por consiguiente, el ajuste de ambos modelos es prácticamente el mismo.

Llegados a esta instancia, es posible realizar una comparación más refinada sobre el uso alternativo de los factores IMC y OC. En lo que respecta a los hombres, la principal diferencia en el uso de estas dos medidas antropométricas radica en el efecto que las mismas ejercen sobre la pre-diabetes (comparar primer gráfica entre Figuras 6 y 8). En concreto se aprecia que el efecto de la OC es ciertamente más oscilante que el del IMC. La oscilación mas relevante es la meseta que se produce entre los 92 y los 100 cm. de OC. Tal meseta

estaría explicada por la existencia de individuos altos, para los cuales poseer una cintura de entre 90 y 100 cm. no representa un síntoma de obesidad. Al mezclarse estos individuos con otros que poseen una OC similar pero de estatura baja (presentando entonces síntomas de obesidad), se produce esta meseta en el efecto en cuestión.

La explicación previa sugiere que el IMC, por su propia construcción, no confunde el sobrepeso con una mayor estatura de las personas. Sin embargo, como medida de obesidad (o sobrepeso) el IMC no tiene en cuenta la influencia que ejerce la complexión física de los individuos sobre el peso corporal. La complexión física se suele clasificar, para ciertos fines, como pequeña, mediana y grande. En lo que respecta a la OC, la misma tampoco tiene en cuenta explícitamente a la complexión física, aunque posiblemente esté menos relacionada con ella en comparación al IMC.

De todo lo comentado, para el grupo de los hombres, se deduce que lo mejor sería poseer una medida directa del sobrepeso, construida como el diferencial entre el peso real registrado y el peso "ideal", calculando este último teniendo en consideración la estatura y la complexión física de la persona. Pero lo habitual es que la complexión física de los individuos no se registre en las encuestas.

Con respecto a las mujeres, no se verifican diferencias marcadas entre el uso del IMC y la OC como covariables (compárese las Figuras 7 y 9). Por este motivo, las consideraciones efectuadas anteriormente para los hombres no son trasladables al contexto de las mujeres. Lo que se aprecia es una leve no linealidad en los efectos del IMC, a diferencia de los efectos de la OC que son perfectamente lineales. Esta diferencia, como es lógico, también se refleja en una diferencia leve de poder explicativo a favor del IMC, la cual se puede apreciar revisando los criterios AIC y GCV de los modelos 5 y 6 que figuran en el Cuadro 2 (Sección 4.3).

Para finalizar esta sección se presenta en las Figuras 10 y 11, para hombres y mujeres respectivamente, los efectos estimados del IMC y la OC cuando ambas covariables se incluyen simultáneamente en el modelo.

Tanto para hombres como para mujeres se observa una reducción de todos los efectos y un aumento en la amplitud de los intervalos de confianza, como era de esperar a priori dada la estrecha relación entre el IMC y la OC.

En el caso de los hombres, solo los efectos del IMC y la OC sobre la PreDM podrían justificar la inclusión conjunta de ambas variables antropométricas, para fines de interpretación sobre todo, ya que tales efectos difieren sustancialmente. Respecto a la influencia sobre la DM, el efecto de la OC es prácticamente nulo, mientras que el del IMC presenta intervalos muy amplios.

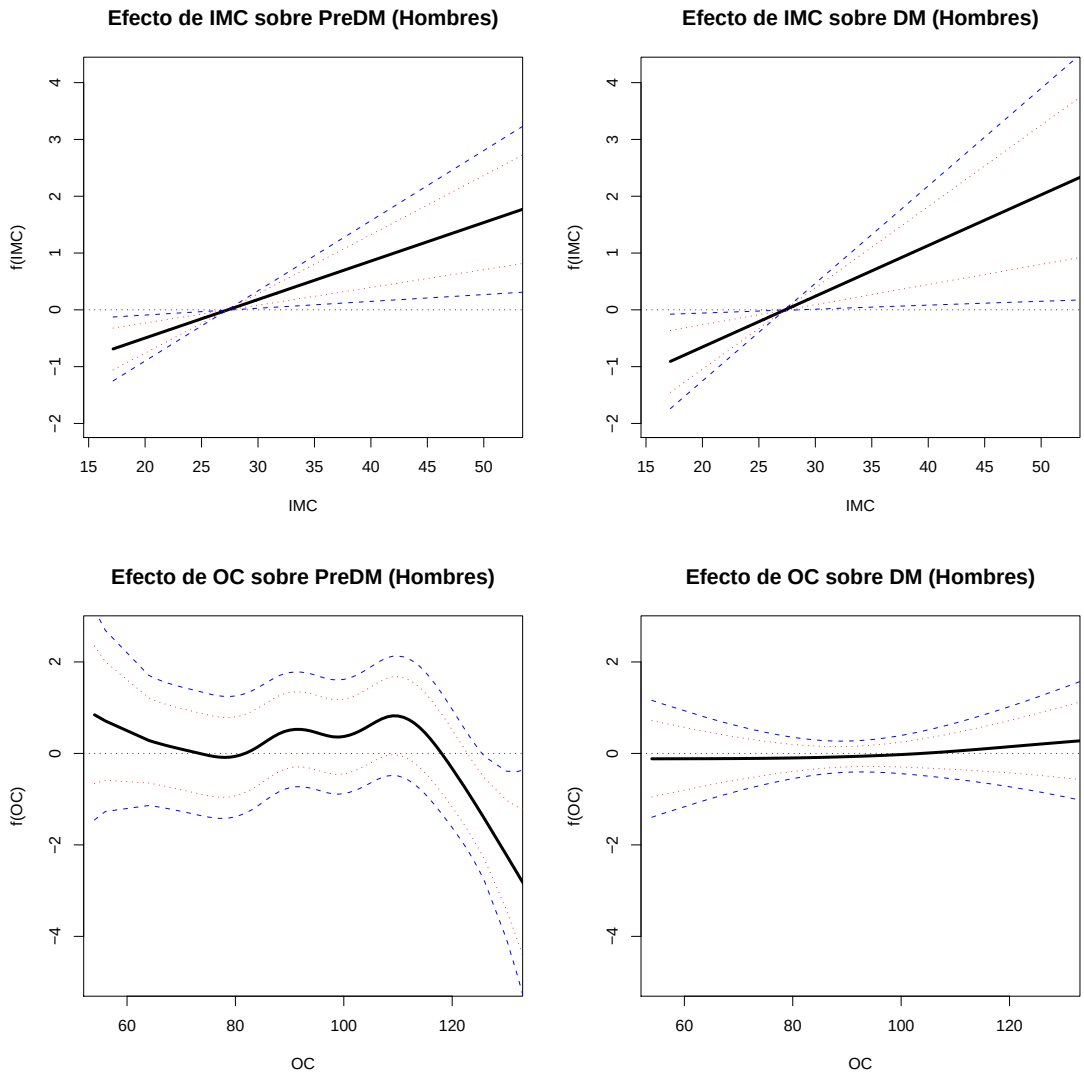

Figura 10: *Efectos centrados del IMC y la OC sobre el log-odds de PreDM y DM, estimados conjuntamente en el mismo modelo y para el caso de los hombres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

Para las mujeres los resultados son algo mas curiosos. Por un lado el IMC es claramente superior (y significativo) que la OC en cuanto al efecto sobre PreDM. Pero, por otro lado, la OC es la variable significativa para explicar la DM.

Estos resultados demuestran, al menos, la necesidad de tomar precauciones a la hora de seleccionar una variable antropométrica representativa del sobrepeso. Esto es particularmente importante si el objeto del estudio es explicar la relación entre la obesidad y las prevalencias

de diabetes y pre-diabetes.

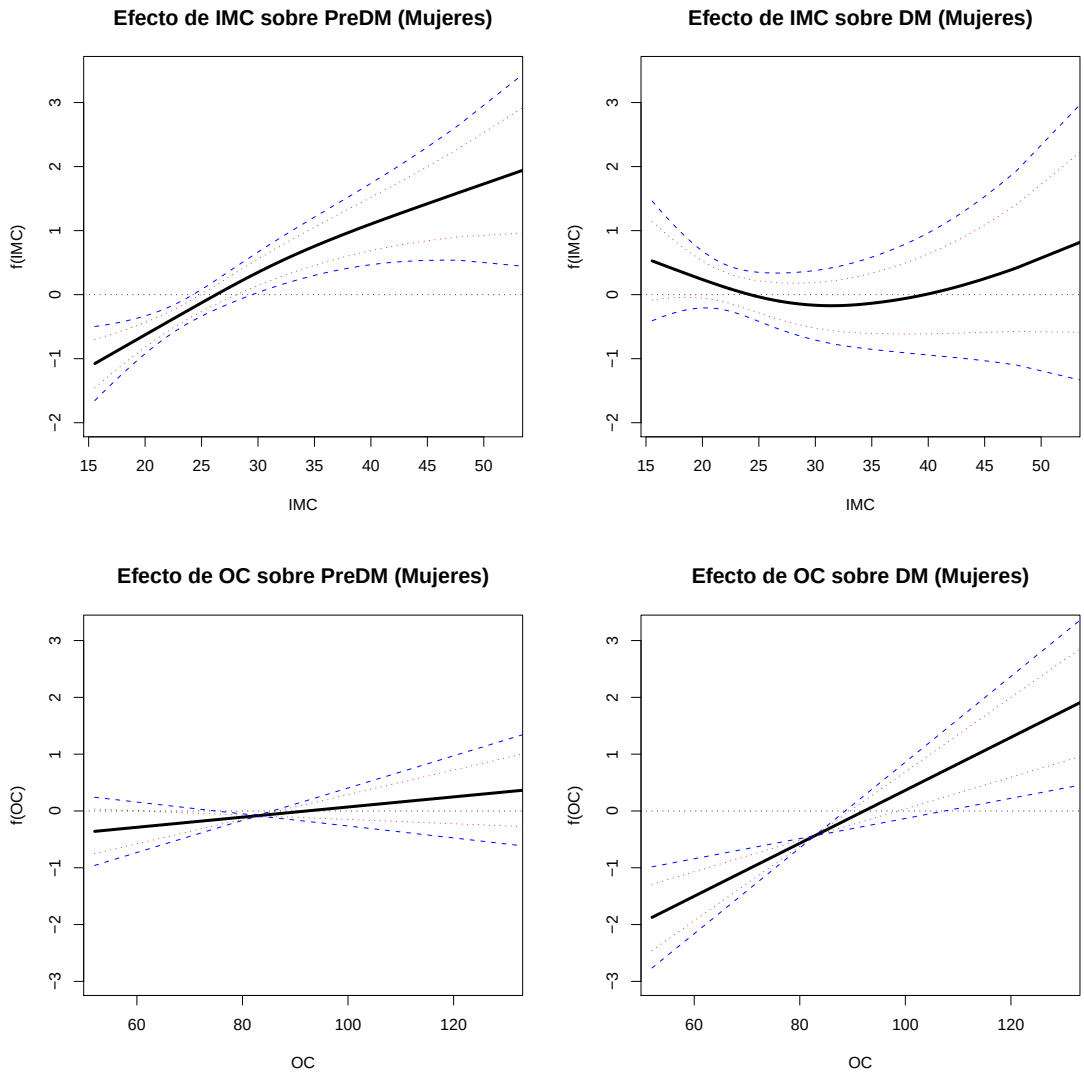

Figura 11: *Efectos centrados del IMC y la OC sobre el log-odds de PreDM y DM, estimados conjuntamente en el mismo modelo y para el caso de las mujeres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

5.2 Modelo Logit Multinomial Acumulativo para explicar la diabetes y pre-diabetes

Anteriormente se mencionó que si bien *gluco* es una variable categórica ordinal, en principio se la trataba como nominal con el fin de seleccionar el estado "normal" como categoría de referencia y poder modelar las prevalencias de pre-diabetes y diabetes en relación a dicho estado de referencia. Tal enfoque es útil para analizar por separado las mencionadas prevalencias, pero no toma en consideración el carácter ordinal de *gluco*.

En esta sección se analiza la variable *gluco* explotando su carácter ordinal, mediante la estimación de un modelo Logit Acumulativo (véanse Secciones 2.3 y 3.1). En este marco, la log-odds *acumulativa* del modelo de probabilidad acumulada viene dada por

$$\log\left\{\frac{P(Y_i \leq r)}{P(Y_i > k)}\right\} = \gamma_{r0} - \gamma_1 educ_i - f_{edad}(edad_i) - f_{IMC}(IMC_i) = \eta_i^r, \tag{43}$$

en donde $r \in \{\text{PreDM, DM}\}$.

La probabilidad $P(Y_i \leq r)$ se define, en forma análoga al caso nominal, como

$$P(Y_i \leq r) = \frac{\exp(\eta_i^r)}{1 + \exp(\eta_i^{PreDM}) + \exp(\eta_i^{DM})}. \tag{44}$$

donde η_i^{PreDM} y η_i^{DM} solo difieren en el intercepto γ_{r0}. Por lo tanto, en esto modelo existe un único efecto "global" de las covariables sobre la probabilidad (o prevalencia) acumulada.

Nótese que, a diferencia de los modelos ordinales descriptos en las Secciones 2.3 y 3.1, el predictor η_i^r en (43) agrega el signo negativo (-) a los efectos de las covariables. Este cambio no afecta al en nada al modelo, solo cambia la forma de interpretar los efectos de las covariables. En este sentido, el cambio ayuda a guardar la lógica seguida hasta el momento, es decir, ver el efecto del IMC y la OC sobre el riesgo de pasar a un estado de mayor nivel de glucosa, es decir, ver el efecto sobre la siguiente log-odds:

$$\log\left\{\frac{P(Y_i > r)}{P(Y_i \leq k)}\right\}. \tag{45}$$

En definitiva, con este modelos se espera que los efectos del IMC y la OC posean tasas de crecimiento positivas.

Nuevamente se efectúan las estimaciones separando a mujeres y hombres. Se utilizan las mismas covariables que en la sección anterior, es decir, la edad, y la educación, además del IMC y la OC. La Figura 12 presenta, para el caso de los hombres, los efectos del IMC y de la OC sobre la log-odds acumulativa (45). Tales efectos fueron estimados en modelos separados,

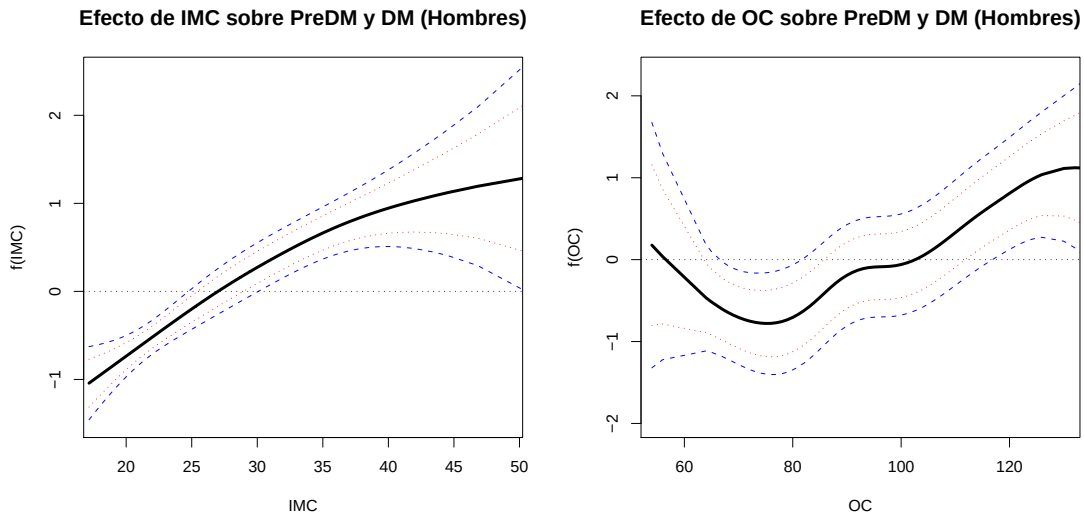

Figura 12: *Efectos centrados del IMC y la OC sobre el log-odds acumulativa, estimados para los hombres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

es decir, un modelo con solo el IMC como variable antropométrica y otro modelo con la OC.

Al igual que para la pre-diabetes en el caso nominal (ver Figuras 6 y 8), en esta oportunidad se verifica una mayor oscilación en el efecto de la OC respecto al del IMC. Tales diferencias se pueden explicar con los mismos argumentos esgrimidos en el marco del modelo de respuesta nominal, expuestos en la sección previa.

Es interesante notar cómo estos efectos ("globales") sobre la log-odds acumulativa parecen combinar o resumir los efectos individuales (sobre PreDM y DM) que se estimaron en el caso nominal.

Para el grupo de mujeres, la Figura 13 contiene las gráficas de los efectos del IMC y la OC sobre la log-odds acumulativa.

No es sorprendente encontrarnos con efectos globales lineales en ambas covariables, dados los efectos que se habían estimado para el caso nominal (ver Figuras 7 y 9).

En lo concerniente a uso alternativo del IMC y de la OC, lo visto hasta el momento no vuelca la balanza para ningún lado. Emulando lo hecho en la sección previa para el modelo nominal, se muestran en las Figuras 14 (para los hombres) y 15 (para las mujeres los efectos

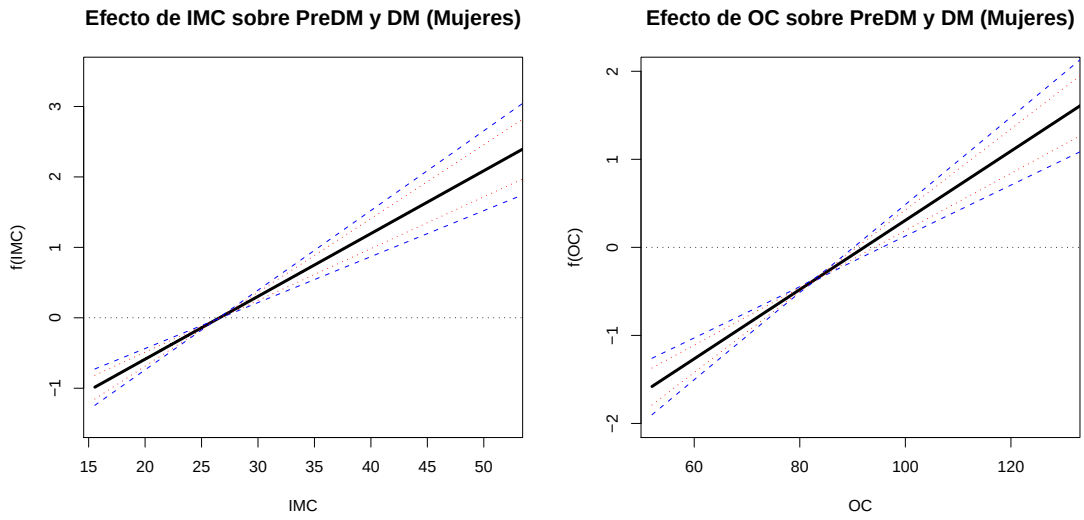

Figura 13: *Efectos centrados del IMC y la OC sobre el log-odds acumulativa, estimados para las mujeres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

estimados al incluir conjuntamente al IMC y la OC en el modelo ordinal.

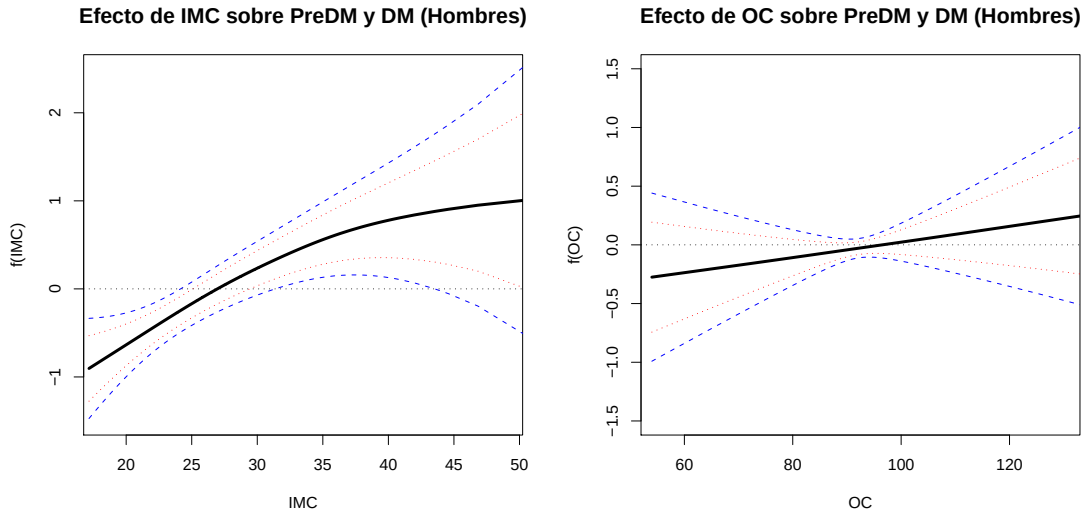

Figura 14: *Efectos centrados del IMC y la OC sobre el log-odds acumulativa, estimados en un único modelo, para los hombres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

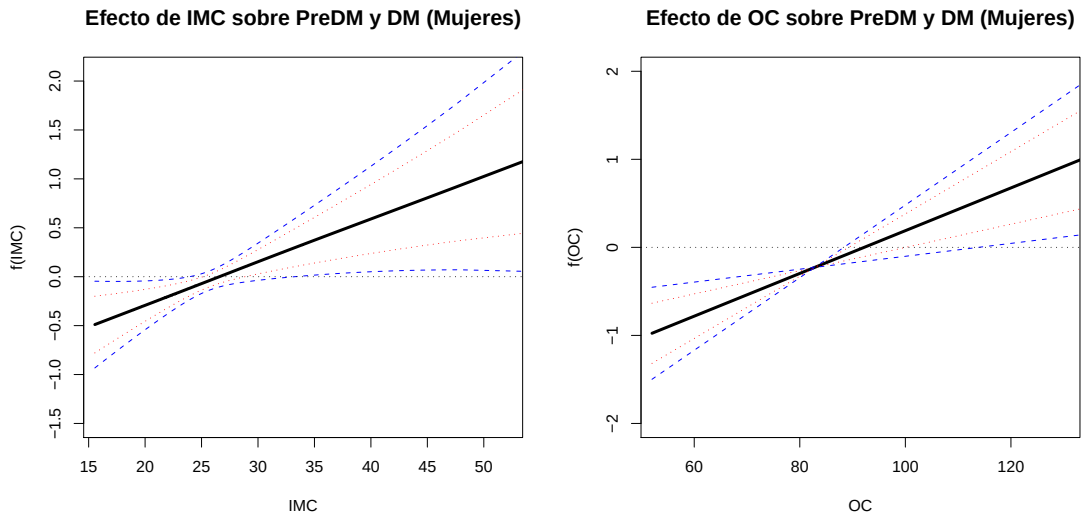

Figura 15: *Efectos centrados del IMC y la OC sobre el log-odds acumulativa, estimados en un único modelo, para las mujeres. Se incluyen los intervalos de confianza del 95 % (trazos discontinuos azules) y del 80 % (trazos punteados rojos).*

Al igual que en el caso de respuesta nominal, la inclusión conjunta de las covariables antropométricas provoca ampliaciones de los intervalos de confianza y reducción de los efectos. En el marco ordinal, lo bueno de tener efectos globales es que se puede determinar más fácilmente si alguna de las covariables eclipsa a la otra o, por el contrario, si ambas variables pueden considerarse en conjunto.

Para los hombres, el IMC parece eclipsar totalmente a la OC, según se aprecia en la Figura 14. Ello sugiere que carece de sentido el uso conjunto de estas covariabes, en todo caso deberían emplearse por separado. En contraste, para el grupo de mujeres no resulta fácil distinguir alguna ventaja a favor de una da las dos covariables consideradas.

Para tratar de echar más luz sobre esta comparación entre IMC y OC, en la sección siguiente se acude al análisis de las curvas ROC de los diferentes modelos con respuesta nominal.

5.3 Comparativa de modelos mediante curvas ROC

Los modelos con respuesta nominal, presentados en las secciones previas, producen estimaciones de las probabilidades $P(Y_i = PreDM)$ y $P(Y_i = DM)$ mediante el uso de la expresión (41). Ambas probabilidades son calculadas respecto de la categoría de referencia

"normal". Por lo tanto es posible evaluar la capacidad de modelos alternativos para clasificar correctamente a los individuos entre dos categorías, en este caso sería la de referencia (normal) y una de las restantes dos (Pre-DM y DM).

Para llevar a cabo la clasificación se requiere fijar un punto de corte c (con $0 < c < 1$), de tal forma que los individuos que posean una probabilidad estimada inferior a ese umbral sean clasificados como normales, mientras que las personas cuya probabilidad estimada supere tal umbral sean consideradas como pre-diabéticas (en caso de clasificar entre normal y Pre-DM) o diabéticas (en caso de discriminar entre normal y DM).

Obviamente, del punto de corte que se fije dependen los errores de clasificación que se cometan. Dichos errores se suelen representar (o cuantificar) mediante dos medidas, la *tasa de falsos positivos* (TFP) y la *tasa de verdaderos positivos* (TVP). La TFP se calcula como el ratio entre la cantidad total de individuos clasificados incorrectamente como "enfermos" (en este caso diabéticos o pre-diabéticos) y el número total de individuos "sanos" (normales). Por su parte, las TVP se define como el ratio entre la cantidad total de personas clasificadas correctamente como "enfermos" y el número total de individuos "enfermos".

Estas dos medidas del error de clasificación dependen del punto de corte, estando inversamente relacionadas, por lo que no es posible reducir una de ellas sin aumentar la otra. Una forma comúnmente usada para visualizar ambas medidas simultáneamente, para todos los valores posibles del umbral c, es a través de la curva ROC. Esta curva representa en el plano a todas las combinaciones de valores TFP y TVP, resultantes de variar el punto de corte c. De esta manera es posible comparar la capacidad de discriminación de varios modelos alternativos.

En la temática que nos concierne, interesa comparar tres modelos diferentes, uno que incluya (además de las covariables edad y educación) solo al IMC como covariable antropométrica, otro que contenga solo a la OC y un tercer modelo que incorpore ambas covariables .

La Figura 16 presenta las curvas ROC, para el grupo de hombres, de los tres modelos mencionados. En la primer gráfica (izquierda) las curvas muestran la capacidad de discriminación entre las categorías "normal" y Pre-DM, mientras que la segunda gráfica (derecha) expone el caso análogo para la clasificación entre categorías "normal" y DM.
Por su parte, la Figura 17 presenta las mismas gráficas que la Figura 16 pero para el caso de las mujeres.

Por lo que se aprecia en ambas figuras, tanto para hombres como para mujeres, las cuatro comparaciones denotan una marcada similitud entre los tres modelos. Pero en los casos "Normal vs. PreDM" para hombres y "Normal vs. DM" para mujeres, el modelo con solo la covariable OC posee ciertas ventajas dentro de zonas importantes. Tales zonas, que

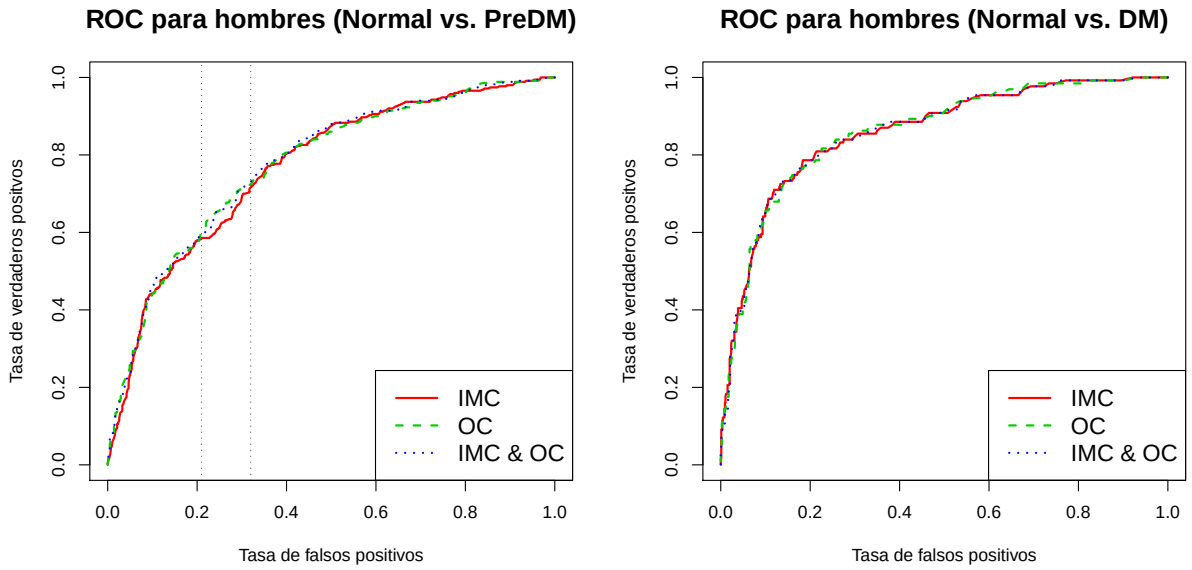

Figura 16: *Curvas ROC para modelos con IMC, con OC y con ambas, estimados para los hombres.*

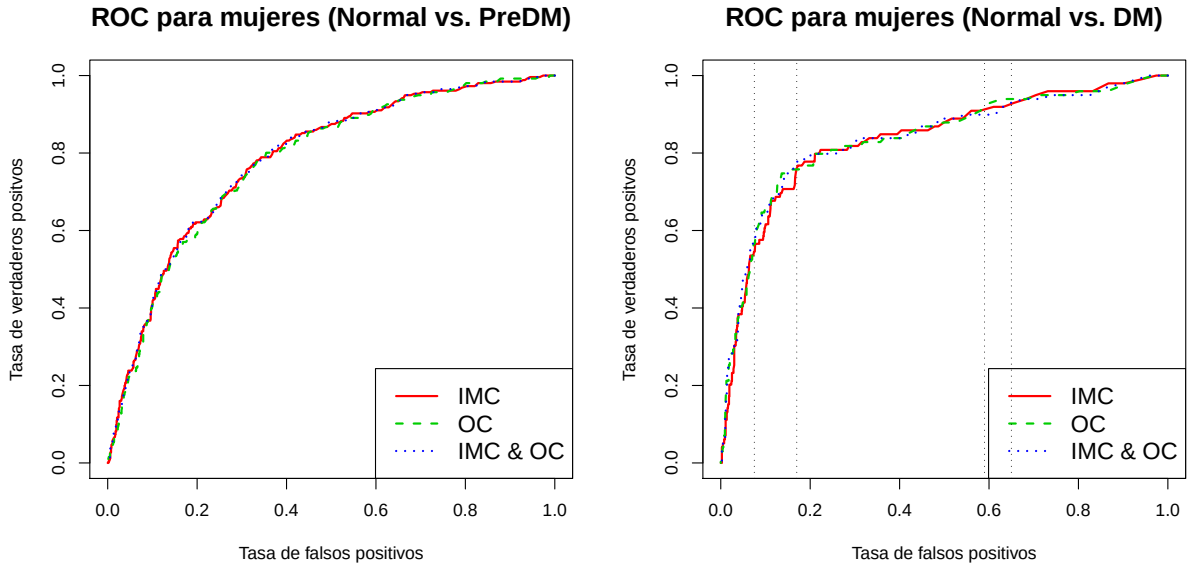

Figura 17: *Curvas ROC para modelos con IMC, con OC y con ambas, estimados para las mujeres.*

se encuentran señaladas en los respectivos gráficos mediante dos líneas punteadas verticales que las contienen, son importantes porque en ellas los valores de la TVP son relativamente

elevados. Las mencionadas ventajas se dan en forma marcada respecto del modelo que solo contiene al IMC. En relación al modelo que incluye tanto la OC como el IMC, tales ventajas son despreciables, como es lógico que suceda.

Además de la comparación visual de las curvas ROC, es común compararlas numéricamente midiendo el área existente entre ellas y una línea recta diagonal que une los vértices inferior-izquierdo y superior-derecho. Dicha área, denominada AUC (*Area Under the ROC Curve*), puede tomar valores entre 0.5 y 1 indicando, a mayor tamaño, un mayor poder de discriminación del modelo. La Tabla 3 presenta los valores de la AUC para los modelos comparados, junto con sus respectivos intervalos de confianza.

Predictor	**Normal vs. PreDM**		**Normal vs. DM**	
$(\eta = \gamma_{r0} + \ldots)$	AUC	IC 95 % boot.	AUC	IC 95 % boot.
Mujeres				
$f(edad) + \gamma_{r1} educ + f(IMC)$	0.7834	(0.7504, 0.8115)	0.8345	(0.7899, 0.8813)
$f(edad) + \gamma_{r1} educ + f(OC)$	0.7781	(0.7492, 0.8080)	0.8369	(0.7974, 0.8841)
$f(edad) + \gamma_{r1} educ + f(IMC) + f(OC)$	0.7843	(0.7555, 0.8171)	0.8375	(0.7888, 0.8832)
Hombres				
$f(edad) + \gamma_{r1} educ + f(IMC)$	0.7680	(0.7400, 0.7972)	0.8597	(0.8239, 0.8915)
$f(edad) + \gamma_{r1} educ + f(OC)$	0.7730	(0.7432, 0.8014)	0.8597	(0.8228, 0.8929)
$f(edad) + \gamma_{r1} educ + f(IMC) + f(OC)$	0.7771	(0.7508, 0.8063)	0.8592	(0.8229, 0.8927)

Cuadro 3: *Valores de la AUC junto con su intervalo de confianza bootstrap (percentil) del 95 %, para los modelos alternativos.*

Los intervalos de confianza para las AUC fueron computados mediante bootstrap (método percentil, con 600 re-muestras), haciendo uso de una rutina programada en lenguaje R (The R-project of Statistical Computing). Solo se aplicó re-muestreo a las probabilidades estimadas con la muestra original, es decir, no se volvió a estimar el modelo de regresión para cada una de las re-muestras. Este es el procedimiento normal cuando la variable de diagnóstico (la probabilidad estimada en este caso) no depende del ajuste de un modelo estadístico. Por este motivo, los intervalos obtenidos son solo aproximaciones "optimistas", ya que la re-estimación de las probabilidades en cada re-muestra incrementaría la amplitud de los mismos.

La inspección de los AUC confirman las apreciaciones efectuadas mediante la comparación visual. En general se percibe una gran similitud entre las AUC de todos los modelos. En particular, la OC tiene una leve ventaja sobre el IMC para el caso de hombres y categorías Normal-PreDM y en el de mujeres y categorías Normal-DM. Mientras que el IMC presenta cierta ventaja sobre la OC para el caso de mujeres y categorías Normal-PreDM. Más allá de las diferencias referidas, la amplitud de los intervalos de confianza de las AUC indican que tales diferencias no son significativas.

También se observa en la tabla que, para la mujeres, los tres modelos discriminan mejor entre las categorías Normal y PreDM respecto de los hombres. La situación inversa se observa para la clasificación entre Normal y DM, para la cual los modelos discriminan mejor en el caso de los hombres. Otra conclusión que se desprende es que, tanto para hombres como mujeres, los tres modelos discriminan mejor entre Normal y DM respecto del caso Normal vs. PreDM.

Considerando tanto el análisis visual como el de las AUC, cabe señalar que el modelo más complejo (el que incluye el IMC junto con la OC) no presenta las ventajas que se podrían esperar a priori, en cuanto a poder de discriminación se refiere.

Si se tuviese que elegir un modelo, teniendo en cuenta todo lo expuesto en esta sección, el elegido sería el modelo con solo la OC como variable antropométrica. Dado este modelo, resulta útil en las investigaciones médicas la comparación del poder de clasificación del mismo entre diferentes escenarios. Para ilustrar estas comparaciones, más allá de los valores de las AUC, en la Figura 18 se presentan las curvas ROC de hombres y mujeres para los dos casos de clasificación estudiados.

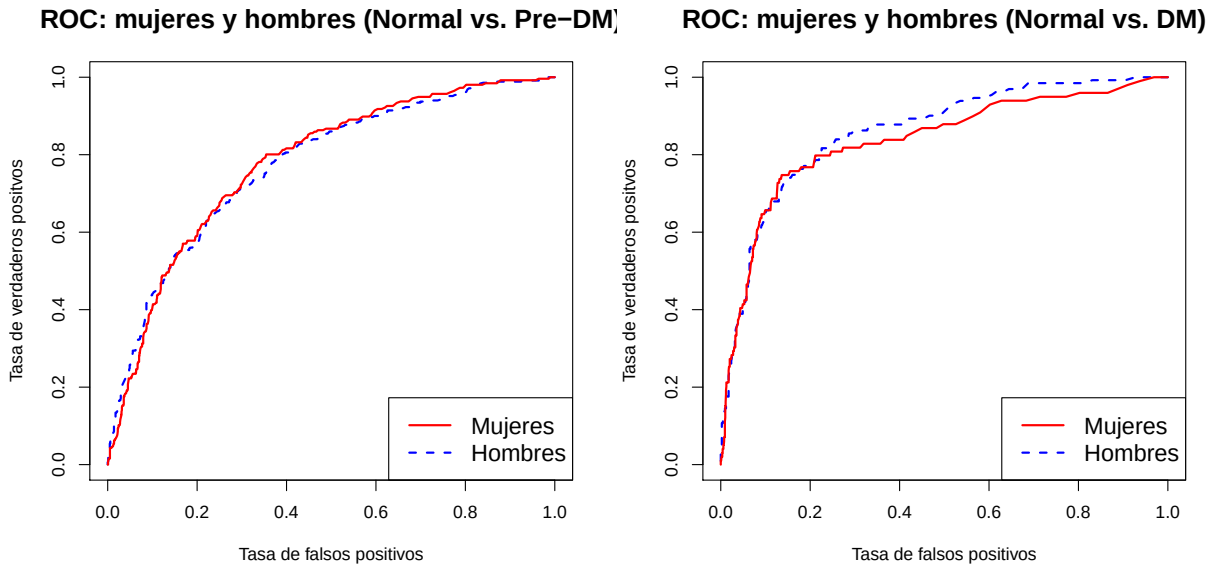

Figura 18: *Curvas ROC para modelo con solo la OC, estimados para mujeres (curvas continuas rojas) y hombres (curvas discontinuas azules).*

En el análisis de los AUC de la Tabla 3 ya se había mencionado que para las mujeres se discriminaba mejor entre Normal y PreDM, mientras que para los hombres se lo hacía mejor entre Normal y DM. Sin embargo, con la comparación de las curvas ROC se tiene una mejor idea de cómo se generan las diferencias. En el caso de la Figura 18, se observa que las

diferencias más sistemáticas se dan a favor de los hombres, cuando la clasificación de interés es entre individuos normales y diabéticos.

De la misma manera, se pueden comparar las curvas ROC del modelo elegido para determinar si es más efectivo discriminando entre individuos normales y pre-diabéticos o entre normales y diabéticos. La Figura 19 presenta las gráficas para tal comparación.

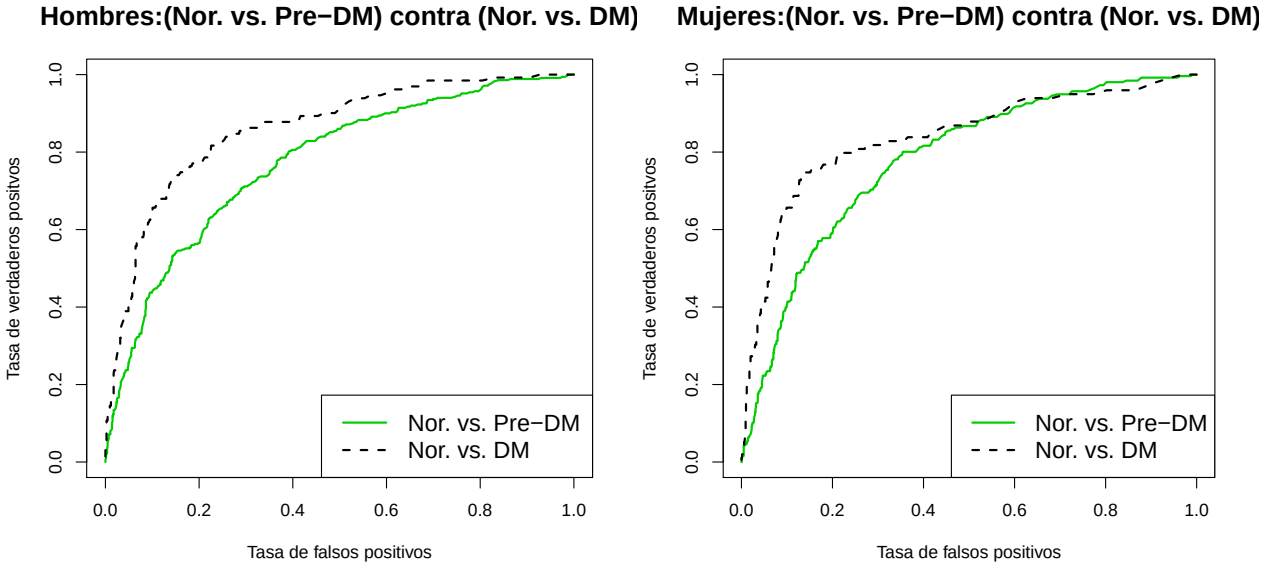

Figura 19: *Curvas ROC para modelo con solo la OC, estimados para mujeres (curvas continuas rojas) y hombres (curvas discontinuas azules).*

Según las AUC de la Tabla 3 el modelo en cuestión discrimina mejor, tanto para los hombres como las mujeres, en el caso de las categorías Normal y DM. Según lo visto en la Figura 19 se puede agregar que para el caso de los hombres las ventajas en dicha discriminación son más sistemáticas, en el sentido que se mantienen consistentemente a lo largo de todo rango de la TFP (tasa de falsos positivos, ubicada en el eje de abscisas).

De esta manera queda ilustrado la utilización complementaria de las curvas ROC y de sus correspondiente áreas AUC. En la literatura especializada existen otras formas de computar las curvas ROC e incluso de contrastar la igualdad entre varias curvas. Tales extensiones metodológicas están más allá del objetivo de este trabajo.

6. Conclusiones

En este capítulo final se resumen las principales conclusiones del trabajo, separándolas en temáticas biomédicas (Sección 6.1) y estadístico-metodológicas (Sección 6.2).

6.1 Conclusiones desde el punto de vista biomédico

De los resultados exhibidos en los Capítulos 4 y 5, se pueden extraer varias conclusiones, las cuales se presentan enumeradas a continuación.

1. La utilización de modelos flexibles como los STAR ha permitido la estimación de efectos complejos de las covariables, en particular para la obesidad central (OC) en el caso de los hombres. El hecho de poder estimar efectos de tal complejidad tiene dos implicancias, una de ellas es la posibilidad de extraer información sobre el comportamiento "local"del efecto de interés. Ello es de utilidad tanto para el contraste de teorías esgrimidas a priori o para la postulación de nuevas teorías. La otra implicancia es la posibilidad de llevar al máximo la capacidad predictiva o discriminatoria de los modelos, lo cual es útil para comparar modelos alternativos pensados con fines de clasificación.

2. Se ha comprobado que las variables IMC y OC son sustitutas a la hora de incluirlas en un modelo que tenga fines de clasificación de individuos entre categorías. En concreto, el modelo más complejo que incluye a ambas covariables antropométricas no se muestra, en general, superior al modelo que solo incluye a la OC, sea en el caso de los hombres como en el de las mujeres. En este sentido, el modelo con solo la OC se evidenció algo superior al que incluye solo el IMC, tanto para la clasificación de hombres entre “normales” y pre-diabéticos, como para la clasificación de mujeres en “normales” y diabéticas. Esto último fue evaluado mediante curvas ROC y sus respectivas AUC, para los modelos con respuesta nominal.

3. Siguiendo con la capacidad discriminatoria, todos los modelos discriminaron mejor entre mujeres normales y pre-diabéticas que entre hombres normales y pre-diabéticos. A su vez, los modelos también se desempeñaron mejor discriminando entre hombres normales y diabéticos respecto de sus análogas mujeres. Las ventajas diferenciales más marcadas se dieron en este último caso.

4. En lo que respecta a la estimación e interpretación de efectos, en modelos de respuesta tanto nominal como ordinal, la covariable OC es la que presenta el efecto mas complejo. De la interpretación de este efecto, y teniendo en cuenta que la OC es la covariable que mejor discrimina entre categorías, se deduce la potencial necesidad de contar con medidas antropométricas adicionales, que permitan cuantificar la complexión (o estructura) física de los individuos. La estructura física, la cual no es medida adecuadamente por el IMC y la OC,

es un factor que influye en el peso de las personas. Por lo tanto, algunas personas de complexión grande que presenten valores elevados de IMC y OC, estarían siendo erróneamente consideradas con sobrepeso o incluso como obesas.

5. Por último, es importante recordar la necesidad abordar el estudio de las diabetes y pre-diabetes haciendo la distinción entre hombres y mujeres, ya que ambos grupos presentan características distintivas propias. Tal distinción es la seguida habitualmente en la literatura médica.

6.2 Conclusiones desde el punto de vista estadístico

Desde el punto de vista metodológico también se obtuvieron conclusiones respecto de algunas limitaciones de la metodología y posibles temas a investigar en el futuro. A continuación enumeramos las principales.

1. Desde el punto de vista de las limitaciones metodológica, el principal aspecto a remarcar es la actual ausencia, en la literatura estadística, de métodos formales para contrastar la especificación de los modelos STAR. Esto nos obligó a recaer en el uso de elementos descriptivos como las curvas ROC y en los criterios de ajuste AIC, BIC y GCV en la comparación de modelos.

2. Otro aspecto sin desarrollar en los modelos STAR concierne a métodos para estimar las derivadas de los efectos. Estas estimaciones son útiles para encontrar puntos de inflección o zonas con mayor o menor influencia de los cambios en las covariables.

3. En la estimación de los modelos STAR se observó que en algunos casos los efectos estimados de la covariable edad eran perfectamente lineales. Ello sugiere que en estos casos se hubiesen podido utilizar componentes paramétricas para especificar el efecto de la edad. Por una cuestión de uniformidad de modelos, en este trabajo se usaron siempre componentes flexibles para el efecto de la edad.

4. Por último se pueden señalar algunos temas pendientes de estudio para futuras investigaciones. Uno de ellos es la inclusión de interacciones entre las variables antropométricas (IMC y OC) y la edad. Esta posibilidad fue explorada para este trabajo, pero presentó problemas de no convergencia de parámetros a estimar junto con un alto costo computacional. Dicho costo se debió al elevado tamaño muestral empleado y al hecho de que la interacción en cuestión implica la estimación de una superficie de regresión.

5. Otro tema es la posibilidad de estimar modelos de respuesta ordinal, con efectos de determinadas covariables específicos para cada categoría, es decir, con efectos que varíen

de categoría en categoría. La especificación de este modelo requeriría un conocimiento mas profundo del fenómeno médico estudiado.

Referencias bibliográficas

- Anjana, M., Saadeep, S., Deepa, R., Vimaleswaran, K.S., Farooq, S. and Mohan, V., S. (2004). Visceral and Central Abdominal Fat and Anthropometry in Relation to Diabetes in Asian Indians. *Diabetes Care*, 27, 2948-2953.

- Belitz, C., Brezger, A., Kneib, T. and Lang, S. (2009). BayesX (version 2.00): Reference Manual.

- Belitz, C., Brezger, A., Kneib, T. and Lang, S. (2009). BayesX (version 2.00): Methodology Manual.

- Belitz, C., Brezger, A., Kneib, T. and Lang, S. (2009). BayesX (version 2.00): Tutorials.

- Brezger, A., Kneib, T. and Lang, S. (2005). BayesX: Analyzing Bayesian Structured Additive Regression Models. *Journal of Statistical Software*, Vol. 14, issue 11.

- Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, 50, 967-991.

- Botana, M.A., Mato, J.A., Cadarso-Suárez, C., Tomé, M.A., Pérez-Fernández, R., Fernández-Mariño, A., Rego-Iraeta, A. and Solache, I., (2007). Overweight, obesity and central obesity prevalences in the region of Galicia in Northwest Spain. *Obesity and Metabolism*, 3, 3, 106-115.

- Fahrmeir L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models. Springer-Verlag*, New York.

- Faeh, D., William, J., Tappy , L., Ravussin, E. and Bovet, P. (2007). Prevalence, awareness and control of diabetes in the Seychelles and relationships whit excess body weight. *BMC Public Health*, 7:163.

- Hastie, T. and R. Tibshirani, (1990). *Generalized Additive Models.* Chapman and Hall. London.

- Hastie, T. and R. Tibshirani, (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society*, Series B, 55, 757-796.

- Jean-Baptiste, E.D., Larco, P., Charles-Larco, N., Vilgrain, C., Simon, D. and Charles, R., (2006). Glucose intolerance and other cardiovascular risk factors in Haiti (PREDIAH). *Diabetes Metab*, 32, 443-451.

- Kim, S.M., Lee, J.S., Lee, J., Na, J.K., Han, J.H., Yoon, D.K., Baik, S.H., Choi, D.S. and Choi, K.M., (2001). Prevalence of Diabetes and Impaired Fasting Glucose in Korea. *Diabetes Care*, 29, 2, 226-231.

- King, H., Auber R.E. and Herman W. H. (1998). Global burden of diabetes, 1995-2025: prevalence, numerical estimates and projections. *Diabetes Care*, 21, 1414-1431.
- Kneib, T., Fahrmeir L. (2006). Structured Additive Regression for Categorical Space-Time Data: A Mixed Model Approach. *Biometrics*, 62, 109-118.
- Lang, S. and Brezger, A. (2004). Generalized structured additive regression based on Bayesian P-splines. *Journal of Computational & Graphical Statistics*, 13, 183-212.
- Marx, B.D. and Eilers, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis* 28, 193-209.
- Yach, D., Stuckler, D., and Brownell, K.D. (2006). Epidemiologic and economic consequences of the global epidemics of obesity and diabetes. *Nat Med*, 12, 62-66.