CARACTERIZACIÓN DE LA DISTRIBUCIÓN ESPACIAL DE UNA ENFERMEDAD COMPLEJA

Prácticas en el Centro de Investigación en Medicina Molecular y Enfermedades Crónicas (CIMUS)

Máster en Técnicas Estadísticas

2011 - 2012

Alumna: Fátima Valadares Fraga

Directoras: Raquel Cruz Guerrero ; Rosa Mª Crujeiras Casais

II

Índice general

Capítulo 1

Introducción

La Esclerosis Múltiple (EM) es una enfermedad crónica del sistema nervioso central. Se trata de una enfermedad compleja, detrás de la cual hay numerosos factores genéticos (principalmente genes relacionados con el sistema de los antígenos leucocitarios humanos (Human Leukocite Antigen, HLA)) pero en la cual también hay una importante influencia ambiental (Handel et al., 2010 [6]). La prevalencia de EM muestra un gradiente latitudinal, de forma que la enfermedad aumenta conforme nos alejamos del Ecuador (Steve Simpson et al., 2011[12]). Diferentes estudios prueban que el riesgo de padecer esclerosis múltiple cae cuando aumenta el nivel de vitamina D (VD) en la sangre (Van der Mie et al., 2007 [15]). Se sabe que la síntesis de VD depende principalmente del grado de exposición a la luz solar, que correlaciona negativamente con la latitud. Por este motivo con frecuencia se asume que la VD es la responsable del gradiente latitudinal observado en la prevalencia de EM. De esta forma, hay publicados numerosos artículos en los que se relaciona directa o indirectamente la prevalencia de EM con los niveles de VD en personas, con la latitud o con la luz ultravioleta o irradiación (Acheson et al., 1960 [1], Sutherland et al., 1962 [13], Van der Mie et al., 2001 [16]).

Pero el papel de la VD sobre la distribución de la EM es en realidad algo más complejo. Por un lado porque los niveles de VD dependen de más factores que la irradiación, como es el caso de la dieta (un mayor consumo en pescado también se ha relacionado con diferencias en prevalencia de EM)(Swank et al., 1952 [14]). Además, la capacidad de síntesis de la VD varía en función de la pigmentación de forma que pieles claras favorecen la síntesis de la VD mientras que las pieles oscuras presentan dificultades para sintetizarla (Jablonski et al., 2000 [7]), entonces a igual irradiación la síntesis será mayor

en una pigmentación menor. A todo esto hay que añadir, lógicamente, la posible existencia de diferencias genéticas entre las poblaciones europeas en genes relacionados directamente con la enfermedad o con la capacidad de síntesis de VD.

En este trabajo se analizará la relación de la prevalencia de EM con varios de estos factores para intentar caracterizar la distribución espacial de su prevalencia a lo largo de Europa y valorar la contribución relativa de las diferentes covariables.

1.1. Objetivos

El objetivo de este proyecto es la obtención de un modelo de predicción del riesgo de EM en Europa, combinando información genética y ambiental, para explicar el patrón geográfico de distribución de la prevalencia de esta enfermedad (Leibowitz et al., 1967 [8]) y valorar la contribución relativa de diferentes variables. Para ello, se considerarán modelos de regresión que permitan modelar la influencia de distintas covariables sobre la prevalencia de EM, teniendo en cuenta además un posible patrón espacial subyacente. Las variables empleadas para la construcción del modelo serán las se mencionan a continuación:

- Variable respuesta: prevalencia de EM.
- Covariables: irradiación, single nucleotide polymorphism (SNP) de pigmentación, latitud y longitud.

La irradiación y la latitud están muy correlacionadas ($r = -0.92$), lo que conlleva a omitir esta última variable en la obtención del modelo. Una vez obtenido el modelo se intentará mejorar introduciendo como variable explicativa la frecuencia de un alelo de riesgo de la prevalencia de EM.

Las variables explicativas pueden presentar dependencia espacial y esto se debe tener en cuenta al hacer el análisis de modelo.

Al tratar de modelar datos con dependencia espacial, suelen distinguirse dos escalas de variabilidad. Por un lado, la variabilidad a gran escala, que marca la tendencia del proceso y que se puede ajustar mediante técnicas de regresión. Por otro lado, la variabilidad a pequeña escala, que describe la estructura de dependencia del error (y por tanto del proceso). El ajuste de este tipo de modelo se puede hacer utilizando iterativas: primero, se estima

la tendencia mediante técnicas de regresión, suponiendo que los datos son independientes. Una vez hecho el ajuste, se construyen los residuos y se analiza si existe estructura de dependencia. Si los residuos son (espacialmente) independientes, el proceso estaría terminado. Por el contrario, si los residuos presentan dependencia espacial, entonces habría que hacer un reajuste al modelo de tendencia, por ejemplo, utilizando ténicas de mínimos cuadrados generalizados. En los casos tratados en este trabajo, una vez ajustada la tendencia, se realizaron tests para comprobar si los residuos eran independientes, sin encontrar indicios de dependencia espacial a pequeña escala, por tanto no fue necesario reajustar los modelos de regresión iniciales.

1.2. Datos

Se valorará la prevalencia de EM, ($prev\_EM$), en diferentes puntos de Europa, y su relación con la distribución geográfica, latitud y longitud, (Lat, $Long$), la irradiación anual (I_anual) y la frecuencia relativa de un SNP representativo de un gen relacionado con la pigmentación (q). Tendremos por tanto:

$$\{(s_i; prev\_EM(s_i); I\_anual(s_i); Lat(s_i); Long(s_i); q(s_i))\} \tag{1.1}$$

donde $\{s_i : i = 1, \ldots, n\}$, son localizaciones de un conjunto de puntos de Europa, siendo n el número de observaciones de la muestra.

La posible influencia de la pigmentación se valora a traves de un SNP (rs16891982) representativo de un gen, SLC45A2, que codifica una proteína transportadora que media la síntesis de melanina. El alelo derivado de este SNP provoca una sustitución no sinónima de una leucina por fenilalanina en la posición 374 del gen, se asocia con piel de color claro y es común sólo en poblaciones de ascendencia europea. El alelo alternativo se asocia a pigmentación oscura y su frecuencia en poblaciones europeas disminuye a medida que aumenta la latitud.

En la base de datos ALFRED (http://alfred.med.yale.edu/alfred/) se recogen datos de frecuencia poblacional para este SNP en distintos puntos de Europa. Esta base de datos es una libre recopilación de datos de frecuencia de los alelos de los polimorfismos de secuencias de ADN en diversas poblaciones humanas. Para este SNP hay datos de frecuencias en 35 muestras de Europa. A esta información se le añadieron datos puntuales obtenidos de otra base de

datos sobre estudios de asociación en melanoma (http://www.melgene.org/) utilizando en este caso la frecuencia del SNP en la muestra control.

Tras promediar los datos correspondientes a poblaciones de los mismos puntos geográficos en los que se pudo obtener información sobre la prevalencia, se obtuvo un total de 23 puntos reflejados en el mapa de la Figura 1.1.

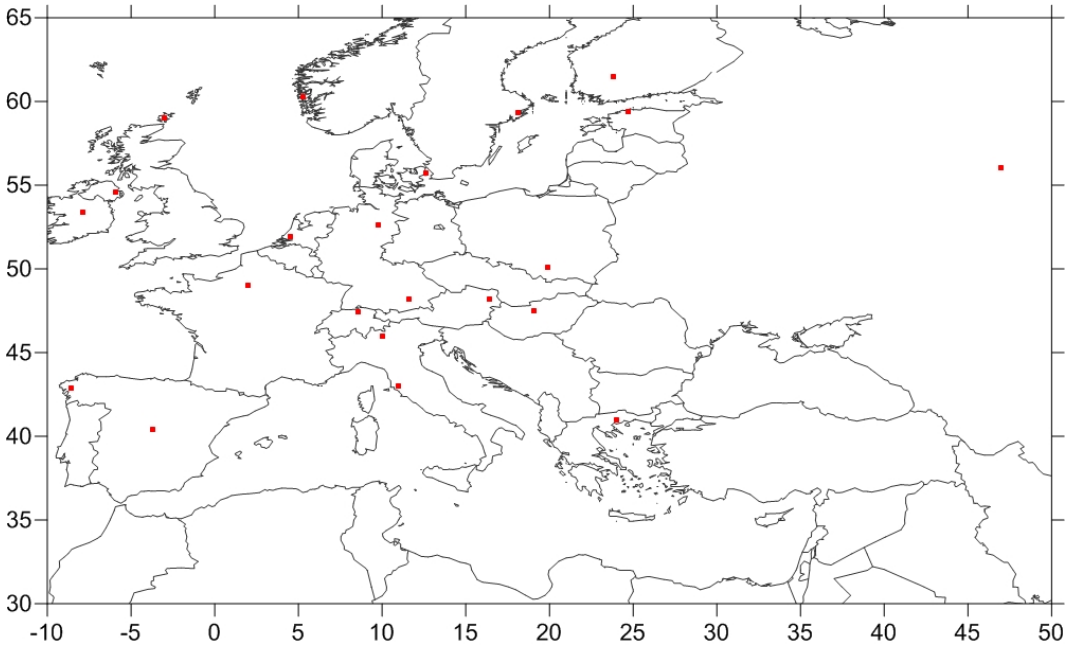

Figura 1.1: Distribución geográfica de los puntos donde se dispone de información sobre las variables de interés.

Los datos de la prevalencia de EM se obtuvieron del artículo de revisión "The epidemiology of multiple sclerosis in Europe"(Pugliatti et al., 2006 [9]).

En esos puntos geográficos se obtuvo una estimación de la irradiación anual a través de la web PVGIS (Photovoltaic Geographical Information System). PVGIS proporciona un registro basado en mapas de recursos de energía solar y en la valoración de la generación de electricidad a partir de sistemas fotovoltaicos en Europa, África y Asia sudoccidental. La base de datos de radiación solar PVGIS fue desarrollada a partir de datos climatológicos homogeneizados para Europa y disponibles en el Atlas e Radiación

Solar Europeo. La base de datos consta de mapas ráster que representan las doce medias mensuales y un promedio anual de las sumas diarias de irradiación global para las superficies horizontales, así como aquellos inclinados en ángulos de 15, 25 y 40 grados. Los mapas representan el promedio 1981-1990 de la irradiación global anual ($kWh/m2$); esta aplicación on line (http://re.jrc.ec.europa.eu/pvgis/apps4/pvest.php) permite exportar un archivo con datos de irradiación para un punto geográfico concreto.

En una segunda fase se incluirá en el modelo la frecuencia de un alelo de riesgo de un gen claramente asociado con la EM ($HLA - DRB1 * 03$) para valorar la contribución relativa de las diferencias genéticas a la heterogeneidad en la distribución. Se analizó la contribución de este alelo por ser un alelo de riesgo relacionado con el gradiente latitudinal de la EM (Handel et al., 2010 [6]).

Los datos de frecuencias de este alelo se obtuvieron de la base de datos de Allele Frequency Net Database (AFND) (Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Gonzalez-Galarza FF, Christmas S, Middleton D and Jones AR Nucleic Acid Research 2011, 39, D913-D919.).

En la Tabla 1.1 se presentan los datos de los que se dispone para hacer el análisis. La tasa de prevalencia es el número de casos que presentan EM en un período determinado expresado por 100000 habitantes. Los datos obtenidos pasan el test de normalidad pero con poco margen por lo que se ha decidido transformarlos a escala logit mediante la función:

$$lp = \log\left(\frac{prev\_EM}{1 - prev\_EM}\right)$$

Se denota por *lp* la prevalencia de EM transformada a escala logit. Además, dado que se aplicarán modelos de regresión lineal, esta transformación en la respuesta garantiza la obtención de predicciones en una escala adecuada.

País	**Latidud**	**Longitud**	**Irradiación**	**q**	**q03**
Italia	40	9	4430	0.37	0.18
España	40.4	-3.7	4500	0.17	0.12
Grecia	41	24	3690	0.01	0.08
España	42.9	-8.6	3710	0.21	0.10
Italia	43	11	3790	0.05	0.11
Italia	46	10	3410	0.10	0.09
Suiza	47.4	8.6	3020	0.06	0.09
Hungría	47.5	19.1	3290	0.03	0.10
Alemania	48.2	11.6	3130	0.04	0.10
Austria	48.2	16.4	3140	0.04	0.11
Francia	49	2	3050	0.14	0.12
Polonia	50.1	19.9	2740	0.02	0.10
Holanda	51.9	4.5	2720	0.03	0.12
Alemania	52.6	9.8	2630	0.09	0.11
Irlanda	53.4	-7.9	2530	0.01	0.17
UK	54.6	-5.9	2520	0.03	0.16
Dinamarca	55.7	12.6	2670	0.03	0.11
Rusia	56	47	2750	0.05	0.07
UK	59	-3	2220	0.02	0.15
Suecia	59.3	18.1	2310	0.25	0.08
Estonia	59.4	24.7	2620	0.01	0.08
Noruega	60.3	5.3	2200	0.02	0.13
Finlandia	61.5	23.8	2480	0.08	0.07

Tabla 1.1: Medidas de latitud, longitud, irradiación anual, frecuencia relativa de SNP de pigmentación (q) y frecuencia del alelo de riesgo HLA-DRB1*03 (q03) en los puntos observados.

Capítulo 2

Materiales y Métodos

En este capítulo se hace una breve introducción al análisis de regresión lineal múltiple y a la construcción de un modelo de regresión, así como la validación del mismo (Faraway, 2004 [4], Sánchez, 2010 [10]).

2.1. Regresión lineal múltiple

Para estudiar la dependencia entre una variable Y (variable repuesta o dependiente) y un conjunto de variables $X_1, \ldots, X_{p-1}$ (variables explicativas o independientes), se utilizan los modelos de regresión.

Estos modelos pueden tener una sola variable explicativa o más. En el caso de tener un conjunto de variables explicativas $(X_1, X_2, \ldots, X_{p-1})$, es decir, si se considera que los valores de la variable dependiente Y son generados por una combinación lineal de más de una variable explicativa, entonces se tiene un modelo de regresión múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon,$$

donde $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son los parámetros que acompañan a las variables y ϵ el error.

Hay ciertos requerimientos necesarios para poder utilizar la técnica de regresión lineal:

- Linealidad: se supone que la variable respuesta depende linealmente de las variables explicativas.

- Homocedasticidad: la varianza de los residuos es constante.

- Normalidad: no es suficiente que el error sea pequeño, sino que además debe seguir una distribución normal, $\epsilon \in N(0, \sigma^2)$.

- Independencia de los errores: bajo la suposición de tener una muestra aleatoria simple generada por el modelo, los errores correspondientes a dichas observaciones son independientes.

Para estimar los parámetros del modelo se necesita una muestra. Se distinguen dos tipos de diseño experimental, en este caso, se utiliza una muestra con diseño aleatorio, esto es, tanto la variable repuesta como las variables explicativas son variables aleatorias. La muestra resultante de un diseño aleatorio es de la forma: $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2.1.1. Coeficiente de correlación

Sean X e Y dos variables aleatorias. Se denomina coeficiente de correlación simple entre X e Y al cociente de la covarianza entre ambas y el producto de las desviaciones típicas, es decir,

$$\rho_{XY} = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

En el caso de disponer de una muestra aleatoria simple de n observaciones de X e Y, el coeficiente de correlación simple se puede definir como

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}, \tag{2.1}$$

donde S_X y S_Y son las desviaciones típicas muestrales de X e Y respectivamente, y S_{XY} denota la covarianza muestral de X e Y.

Este coeficiente puede tomar valores desde menos uno hasta uno, $-1 < \rho < 1$, indicando que mientras más cercano a uno sea el valor del coeficiente de correlación, en cualquier dirección, más fuerte será la asociación lineal entre las dos variables. Cuanto más cercano a cero sea el coeficiente de correlación, este indicará que más débil es la asociación entre las variables.

2.1.2. Interacciones

Dos variables interaccionan cuando el efecto de una de ellas sobre la respuesta depende del nivel de la otra.

La forma en la que habitualmente se afronta el problema de determinar si existe un efecto de interacción entre las variables explicativas es mediante la inclusión en el modelo teórico de los productos de algunas de las variables explicativas. Por tanto, en el caso de un modelo con dos variables explicativas, estableceríamos el modelo con interacción del siguiente modo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon.$$

En el modelo con interacciones, β_0 se denomina constante o término independiente, los coeficientes β_1 y β_2, que acompañan a las variables explicativas, se les denomina coeficientes asociados a los efectos principales, y a β_{12} se conoce como coeficiente de interacción.

2.1.3. Selección de variables

Los métodos de selección permite especificar cómo se introducen las variables independientes en un modelo de regresión. Utilizando distintos métodos se pueden construir diversos modelos de regresión a partir del mismo conjunto de variables. El objetivo de estos métodos es encontrar el modelo que mejor ajuste los datos y a la vez, que sea lo más sencillo posible. Por lo tanto, el objetivo que debe plantearse todo método de selección de variables debe ser el detectar todas aquellas variables que son irrelevantes y/o redundantes en el ajuste.

El procedimiento que se aplicará en este caso es el denominado método backward o método de eliminación hacia atrás. Se comienza con el modelo completo, es decir, el que incorpora todos los efectos que pueden llegar a influir en la variable respuesta, y en cada paso se va eliminando una variable hasta que se considera que no procede eliminar ningún términos más.

Para realizar este procedimiento, contamos en R con la función *step* que selecciona el modelo a partir del criterio de información de Akaike (AIC).

El AIC toma en consideración tanto el ajuste del modelo como el número de parámetros utilizados en el mismo, construyendo así una medida global del modelo. Se busca el modelo que tenga el mínimo AIC. Este criterio se define del siguiente modo:

$$AIC = -2\log(L(\beta)) + 2p, \tag{2.2}$$

donde p es el número de parámetros del modelo y $L(\beta)$ la función de verosimilitud asociada al modelo.

2.2. Análisis de los residuos

2.2.1. Validación del modelo

Cuando estudiamos la dependencia de una variable respuesta Y respecto a un conjunto de variables explicativas mediante un modelo de regresión, dicho modelo es elegido entre varios modelos alternativos, ya que los modelos de la regresión sólo son fiables si el modelo cumple ciertas hipótesis sobre los residuos. Es por eso por lo que es necesario realizar una validación del modelo.

El ajuste de un modelo de regresión se puede medir a través del coeficiente de determinación. Dicho coeficiente es la proporción de varianza de la variable dependiente explicada por todas las variables independientes incluidas en el modelo, y se puede calcular como sigue:

$$r^2 = 1 - \frac{RSS}{TSS}, \tag{2.3}$$

donde RSS es la suma residual de cuadrados, $RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ con $\hat{Y}_i$ las predicciones en base al modelo ajustado, y TSS es la suma total de cuadrados, $TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ con $\bar{Y}$ la media muestral de las observaciones de la variable respuesta.

El coeficiente de determinación es el cuadrado del coeficiente de correlación (2.1). Además, éste varía entre cero y uno, indicando que cuanto más próximo esté de uno, mejor explican las variables explicativas la variación de la variable dependiente.

Por tanto, el coeficiente de determinación se puede interpretar como una medida de la bondad de ajuste del modelo de regresión hallado siempre que las hipótesis básicas se cumplan. Si queremos utilizar r^2 para comparar distintos modelos, estos deben tener la misma variable dependiente ya que así tendrán igual suma de cuadrados total. Aún así, esta medida adolece del problema de aumentar su valor al añadir una nueva variable explicativa, sea cual sea su aportación al modelo. Además no tiene en cuenta que hay que estimar un nuevo parámetro con el mismo número de observaciones. Para tener en cuenta este problema se suele utilizar el $r^2 ajustado$ o corregido por grados de libertad. Éste se define como sigue:

$$r^2 ajustado = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}, \tag{2.4}$$

siendo p el número de parámetros de regresión estimados.

Este coeficiente será mejor cuanto más alto sea su valor y puede disminuir si se introducen variables cuya incorporación en el modelo no sea significativa.

2.2.2. Diagnosis del modelo

En los modelos de regresión se suponen unas hipótesis básicas que debemos comprobrar si realmente se verifican. Es lo que se conoce como diagnosis del modelo.

La diagnosis del modelo se puede realizar mediante el análisis de los residuos. Aunque las hipótesis de homocedasticidad, normalidad e independencia se refieren a los errores, éstos se aproximan por los residuos. Los residuos no coinciden uno a uno con los errores correspondientes, pero el comportamiento del conjunto de los residuos sigue el comportamiento del conjunto de los errores, y por tanto seguirá el cumplimiento o no de las hipótesis.

Linealidad

La variable dependiente es función lineal del conjunto de variables explicativas. Para comprobar esta hipótesis, se utilizará un diagrama de dispersión de los valores ajustados sobre los observados.

Normalidad

Existen diferentes test estadísticos para verificar la normalidad. Entre ellos está el test de Shapiro-Wilk, en el que se supone como hipótesis nula que la muestra proviene de una población normalmente distribuida.

Heterocedasticidad

Un modelo heterocedástico es aquel en que las varianzas de los errores no son constantes, por lo tanto, la variabilidad es diferente para distintos valores de la variable explicativa.

Independencia

Una hipótesis básica en el estudio de los modelos de regresión lineal es la independencia de los errores asociados a cada observación, esto es, los errores $\{\epsilon_i\}_{i=1}^{n}$ son variables aleatorias independientes.

Hay situaciones en las que las observaciones se toman en individuos diferentes y por tanto no hay por qué sospechar una correlación entre ellos. En estos casos, es lógico suponer que los errores son independientes. Sin embargo, hay otras situaciones en las que es posible que exista correlación. Aunque hay varias formas de correlación, la más usual es autocorrelación o correlación serial. Esta correlación surge cuando se toman observaciones sobre una misma situcaión en diferentes instantes de tiempo.

En los datos que se tienen para este estudio se intuye que puede existir otro tipo de correlación, en contreto dependencia espacial, por tanto, se deben utilizar otro tipo de técnicas para verificar que exite independencia. Este tipo de técnicas son propias de la geoestadística, una rama de la estadística que trata fenómenos espaciales (Diggle y Ribeiro 2007 [3], Schabenberger y Gotway 2005 [11]). Su interés primordial es la estimación, predicción y simulación de dichos fenómenos. Consideremos proceso en el espacio $\{Z(s) : s \in D \subset \mathbb{R}^d\}$ observado en ciertos puntos $\{s_i : i = 1, \ldots, n\}$. Una de las hipótesis usuales en estadística espacial es la estacionariedad intrínseca que implica:

a) $E(Z(s) - Z(s+u)) = 0$.

b) $Var(Z(s) - Z(s+u)) = E[Z(s+u) - Z(s)]^2 = 2\gamma(u)$,

donde $2\gamma(u)$ es el variograma del proceso. Suponiendo estacionariedad intrínseca del proceso, entonces la estructura de dependencia del proceso espacial quedará especificada por el variograma.

Además, un proceso espacial es isotrópico si la dependencia espacial es solo una función de una distancia escalar, no de la dirección, es decir, si la correlación entre los datos no depende de la dirección en la que esta se calcule. En este caso, $2\gamma(u) = 2\gamma(\|u\|) = 2\gamma(u)$, donde $\|.\|$ denota la norma euclidiana.

Suponiendo estacionariedad intrínseca e isotropía, se define el variograma empírico isotrópico como sigue:

$$2\hat{\gamma}(u) = \frac{1}{|N(u)|} \sum_{(i,j) \in N(u)} (Z(s_i) - Z(s_j))^2$$

donde $N(u) = \{(i,j), \|s_i - s_j\| \sim u\}$ y $|N(u)|$ denota el cardinal de $N(u)$.

El estimador anterior puede verse afectado por observaciones atípicas, por lo que resulta útil introducir un variograma en escala robusta:

$$2\gamma'(u) = \frac{1}{|N(u)|} \sum_{(i,j)\in N(u)} \sqrt{|Z(s_i) - Z(s_j)|} \tag{2.5}$$

El variograma (con su estacionariedad en escala robusta) será una función plana, en el caso de que el proceso sea independiente, o por el contrario, será una función creciente, en el caso de ser dependiente.

Diblasi y Bowman, 2001 [2], propusieron un test para contrastar la independencia en datos espaciales. El contraste se basa en una suavización de la nube de puntos generada por el variograma en la escala robusta.

Este contraste de independencia se aplicará sobre los residuos del modelo de regresión ajustado. Es decir, el proceso espacial Z será en nuestro caso el proceso de los errores ϵ, pero dado que no disponemos de observaciones directas del proceso, se trabajará sobre los residuos.

Colinealidad

Un modelo de regresión presenta un problema de colinealidad cuando las variables explicativas presentan mucha correlación entre sí y, como consecuencia, se consigue explicar la variable dependiente pero no se sabe cuál es el efecto de cada una de las variables explicativas. En general, este problema incrementa la varianza de los estimadores. La solución es eliminar del modelo aquellas variables explicativas que dependen unas de otras. Los métodos de selección de variables solucionan automáticamente este problema.

Una forma de medir la colinealidad es calculando el factor de inflación de la varianza (FIV), definido como $\frac{1}{1-r_j^2}$, siendo r_j^2 el coeficiente de determinación de la $j-$ésima variable explicativa al hacer regresión de ella sobre las demás variables explicativas.

Cuando $r_j^2 = 0$ no existe colinealidad, las variables independientes están incorrelacionadas y su FIV es igual a 1. A medida que el valor de r_j^2 se incrementa en valor absoluto, es decir, existe una correlación negativa o positiva entre las variables, el FIV también se incrementa, ya que el denominador tiende a cero a medida que r_j^2 tiende a uno. Se recomiendan que los FIV sean menores a 5, de lo contrario se concluye que existe colinealidad.

Observaciones atípicas o influyentes

En un modelo de regresión, pueden surgir dos tipos de problemas con algunas de sus observaciones, éstas pueden ser atípicas o influyentes.

Una observación atípica es aquella que está muy separada del comportamiento del modelo de regresión y hace sospechar que no sigue el modelo. Este tipo observaciones se pueden detectar analizando los residuos de la regresión. Residuos grandes apuntan que estas observaciones no siguen el modelo.

Una observación influyente es aquella que tiene un peso muy grande en los coeficientes del modelo, es decir, altera el ajuste del modelo. Una medida global de la influencia de cada observación es la distancia de Cook, que se define del siguiente modo:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y_{j(i)}})}{p\hat{\sigma}^2}, \tag{2.6}$$

con $\hat{Y}_j$ la predicción del modelo de regresión completo para la observación j, $\hat{Y}_{j(i)}$ la predicción para la observación j del modelo ajustado con todos los datos excepto el dato $i-$ésimo y $\hat{\sigma}$ es la estimación de la varianza del error.

Una distancia de Cook grande indica que esa observación tiene un peso considerable en la estimación del modelo.

Capítulo 3

Resultados

En este capítulo se interpretan los resultados de tres modelos de regresión diferentes. En una primera sección el modelo es el ajustado a los datos sin considerar la frecuencia del alelo $HLA - DRB1 * 03$. Este alelo será considerado en la segunda sección. En la tercera y última sección, se consideran todos los datos pero se elimina la observación correspondiente al dato de Cerdeña por ser un valor atípico en relación a la irradiación y prevalencia de EM.

3.1. Modelos de regresión lineal con interacciones

Como se ha citado anteriormente, en un modelo que explique la prevalencia de EM en escala logit (Y), se han considerado como variables respuesta para la construcción de dicho modelo la irradiación ($X_1 = I\_anual$), el SNP de pigmentación ($X_2 = q$) y la longitud ($X_3 = Long$). Tenemos entonces un conjunto de la forma (1.1), ya que las observaciones se han tomado en los puntos indicados en la Figura 1.1.

El proceso de selección que se ha seguido para escoger las variables que finalmente se han introducido en el modelo de regresión es el que se describe a continuación.

En una primera aproximación, se plantea la inclusión de la pigmentación en el modelo, a través de la frecuencia de un SNP representativo, que puede mejorar la caracterización de la prevalencia en función de la irradiación, estudiada en diversos artículos. Se considera además la coordenada geográfica

longitud.

Se introducen en el modelo, además de las variables principales (I_anual, q y $Long$), todas las posibles interacciones entre ellas y se hace una selección de variables con la función *step*. Esta función elimina las variables del modelo cuya presencia no mejora el AIC (2.2) del mismo. El modelo seleccionado es aquel cuyas variables explicativas son las siguientes:

- Irradiación: I_anual
- SNP de pigmentación: q
- Longitud: $Long$
- Interacción de la irradiación y el SNP de pigmentación: $I\_anual * q$
- Interacción de la longitud y el SNP de pigmentación: $q * Long$

Se ha calculado el coeficiente correlación (2.1) entre las variables principales del modelo, obteniendo la siguiente matriz de correlaciones:

$$A = \begin{pmatrix} & lp & I\_anual & Long & q \\ lp & 1.00 & -0.47 & -0.38 & -0.16 \\ I\_anual & -0.47 & 1.00 & -0.12 & 0.76 \\ Long & -0.38 & -0.12 & 1.00 & -0.17 \\ q & -0.16 & 0.76 & -0.17 & 1.00 \end{pmatrix}$$

En la Tabla 3.1 se presentan los resultados del modelo seleccionado.

R^2: 0.62 / R^2 ajustado: 0.51 / AIC: 28.45			
Variables	Estimación	Error Típico	Significación
$Intercept$	-4.46	0.63	1.85e-06
I_anual	-7.93e-04	2.24e-04	2e-03
q	-12.19	7.062	0.10
$Long$	-0.02	0.01	0.01
$I\_anual * q$	3.26e-03	1.75e-03	0.07
$q * Long$	0.17	0.12	0.16

Tabla 3.1: Valores de la estimación de los coeficientes del modelo, del error típico de la estimación y de la significación de los coeficientes.

A partir de los resultados de la Tabla 3.1, se observa que a mayor irradiación anual, menor prevalencia. Como se ha comentado anteriormente, la variable de irradiación está altamente correlacionada con la latitud y para evitar el problema de la alta colinealidad entre variables, la variable latitud se excluyó del modelo de regresión. La pendiente negativa de irradiación en el análisis de regresión múltiple encaja con lo publicado, resume el gradiente latitudinal observado (al no meter latitud en el modelo, su efecto se refleja en esta variable) y apunta indirectamente a un papel de la VD en la EM (a menor radiación, menor síntesis y mayor riesgo).

La pendiente negativa de la longitud también refleja un gradiente observado en EM aunque generalmente menos comentado, a menor longitud mayor prevalencia, lo que indica que hay mayor prevalencia en el oeste (menor longitud) que en el este de Europa. Esta relación no llega a ser significativa en un análisis univariante ($r = -0.35$, $p - valor = 0.10$) pero se pone de manifiesto al tener en cuenta la irradiación (y la frecuencia del SNP de pigmentación, q). En conjunto, entre estas dos variables, ilustran un gradiente más bien sureste-noroeste (SE-NW), como se ve en el mapa de la Figura 3.1.

La pendiente negativa de q -aunque no llega a ser significativa- sería opuesta a lo esperado según la hipótesis inicial. Esta variable es la frecuencia del alelo de menor frecuencia, que en este caso indica pigmentación oscura, por lo que una pendiente negativa indicaría una mayor prevalencia en poblaciones con piel clara, supuestamente mejor preparadas para la síntesis de VD. Realmente, este resultado es lógico, ya que también hay un claro gradiente latitudinal para pigmentación. De hecho, la distribución de este rasgo se considera un claro ejemplo de la actuación de la selección natural en humanos, favoreciendo pieles claras en zonas de baja irradiación para optimizar la síntesis de VD y evitar raquitismo y demás. Para encontrar el efecto -quizás demasiado sutil- sobre la prevalencia de EM habría que tener una mayor densidad de puntos.

La interacción irradiación-pigmentación ($I\_anual * q$) -aunque no llega a ser significativa, pero por muy poco- presenta una pendiente positiva, lo que significaría que el efecto de la irradiación sobre la prevalencia aumenta a medida que aumenta el valor de q. Se tendría así que en el sur, donde la frequencia del alelo de pigmentación oscura es más alta, a mayor irradiación mayor prevalencia, pero esto puede ser debido al efecto del dato de Cerdeña, una población con prevalencia excepcionalmente alta y alta irradiación documentada en la bibliografía. Análogamente, el efecto de la q (también negativo, aunque no llega a ser significativo) aumenta en función de la irradiación. En

este caso, el efecto de la q aumenta con los valores de irradiación a partir de 3730, que se alcanzan en las poblaciones del sur (Grecia, España e Italia). En estas zonas del sur, a mayor pigmentación mayor prevalencia, pero -del mismo modo que lo que comentamos antes y a pesar de que ahora apoyaría nuestra hipótesis inicial- este efecto se explica exclusivamente por el dato de Cerdeña.

En la Figura 3.1 se ilustra la superficie de tendencia de los valores predichos de la prevalencia de EM para el modelo propuesto de la Tabla 3.1. Como se puede observar, se refleja el gradiente SE-NW que se constata en los resultados.

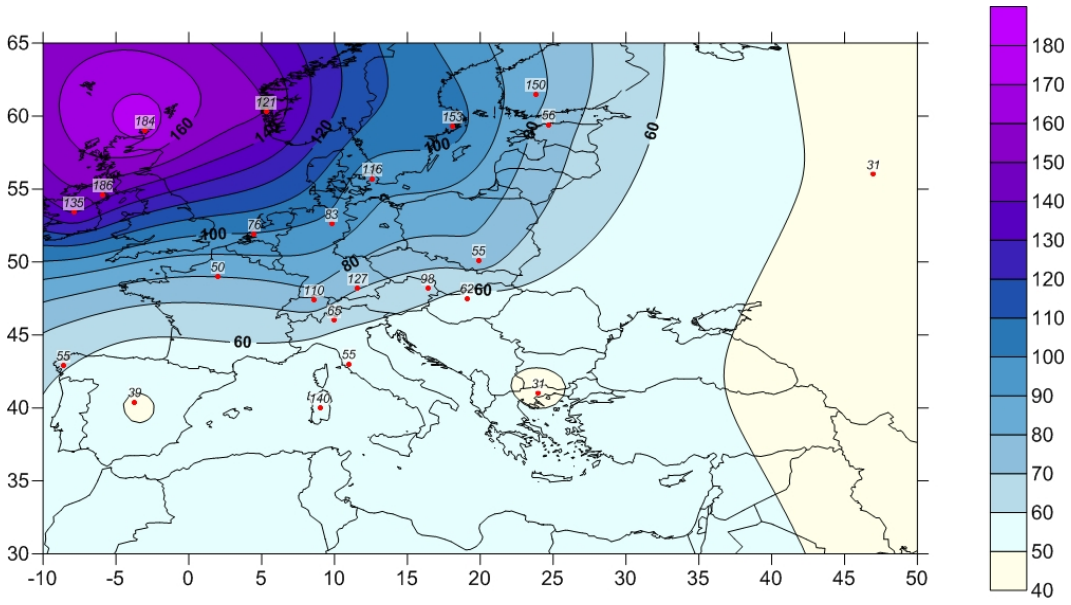

Figura 3.1: Superficie de tendencia de la prevalencia de EM con el modelo estimado. Superposición de los valores observados.

Se debe comprobar si las hipótesis básicas del modelo se dan en los datos considerados este análisis, procedamos por tanto a la diagnosis del modelo.

- Linealidad
  El diagrama de dispersión de la prevalencia de EM frente a los valores

ajustados que se expone abajo (Figura 3.2), muestra que los valores de Y crecen en torno a una línea recta conforme lo hacen los valores ajustados, lo que significa que se verifica la linealidad del modelo.

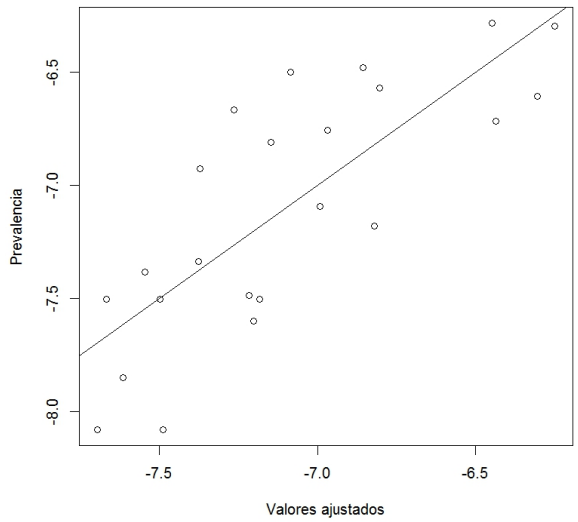

Figura 3.2: Diagrama de dispersión de los valores observados de la prevalencia de EM frente a los valores ajustados del modelo.

- Normalidad de los residuos
  En un modelo de regresión lineal que sea adecuado los residuos deben seguir una distribución normal. Tras aplicar el test de Shapiro-Wilk a los residuos del modelo, se obtiene un p-valor de 0.39, lo que indica que se acepta la normalidad de los mismos.

- Independencia de los residuos
  La independencia espacial de los residuos se evaluó con el variograma (2.5). De la forma plana que presenta el ajuste suavizado del variograma (línea continua en la Figura 3.3) se deduce que no hay dependencia espacial en los errores del modelo. Además, al aplicar el test de independencia de Diblasi y Bowman (Diblasi y Bowman, 2001 [2]) se obtiene un p-valor de 0.36, lo que confirma que no hay evidencias suficientes

para rechazar la hipótesis de independencia. Por tanto, la estructura de los datos quedaría explicada por el modelo de regresión lineal múltiple que se ha ajustado.

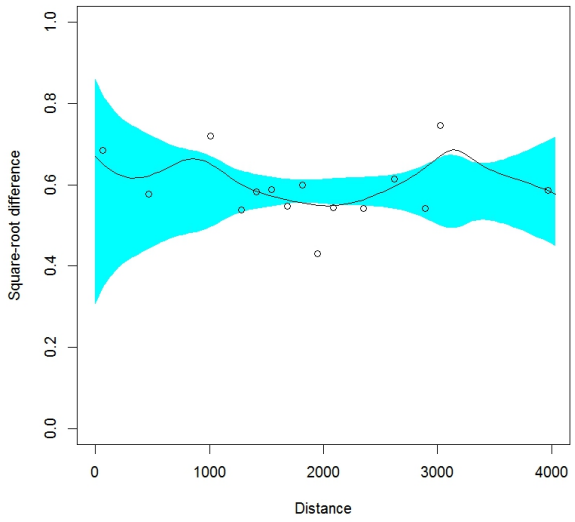

Figura 3.3: Nube de puntos del variograma de los residuos del modelo ajustado en escala robusto. Azul: banda de variabilidad bajo la hipótesis de independencia. Línea continua: variograma suavizado.

- Homocedasticidad
  Cuando no se cumple la hipótesis de homocedasticidad, se dice que existe heterocedasticidad. Una herramienta para detectar la heterocedasticidad es analizar el gráfico de los residuos (véase Figura 3.4).

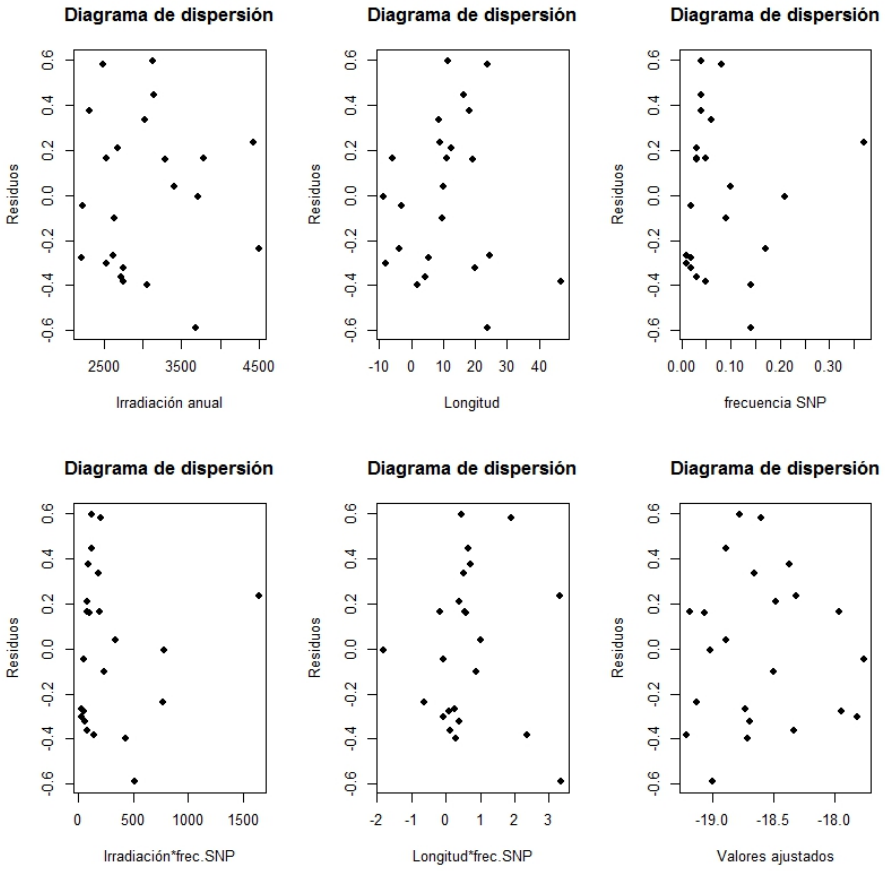

Figura 3.4: Gráficos de los residuos estandarizados frente a cada covariable.

No se refleja de una forma clara la presencia de homocedasticidad y esto se debe a los pocos datos de los que disponemos.

- Observaciones atípicas e influyentes
  Los residuos brutos y estandarizados del modelo, toman valores en (-1,1) y (-2,2), respectivamente, por tanto, no se parece haber ningún dato atípico en la muestra tomada para este análisis.

  Para la detección de observaciones influyentes se han calculado las distancias de cook (2.6). Se consideran preocupantes distancias de Cook a partir de 0.5 o de 1. Sólo supera este umbral, y por tanto altera el

ajuste del modelo, la observación que corresponde al dato de Cerdeña ($D = 6.02$), lo que corrobora que esta observación está interfiriendo en la interpretación de los resultados.

Además de la diagnosis del modelo, se debe hacer la validación del mismo (Subsección 2.2.1). Se ha obtenido un coeficiente de determinación (2.3) de 0.62, lo que indica que el 62 % de la variación de la prevalencia de EM es explicada por las variables explicativas del modelo. El valor del coeficiente de determinación ajustado (2.4) es de 0.51.

3.2. Construcción de la nueva covariable

Una vez obtenido el modelo intentamos mejorarlo incluyendo una nueva covariable, $q03$. Esta covariable es la frecuencia un alelo de riesgo relacionado con el gradiente latitudinal de la EM, alelo $HLA-DRB1*03$. Los puntos de observación que se tienen para este alelo no coinciden con los de la variable respuesta. Además, si se observa el variograma (Figura 3.5) de esta nueva covariable, puede afirmarse que ésta presenta dependencia espacial.

Debemos entonces predecir los valores de la nueva covariable en los puntos donde está medida la variable dependiente.

Cuando el objetivo es hacer predicción, la geoestadística opera básicamente en dos etapas. La primera es el análisis estructural, en la cual se describe la correlación entre puntos en el espacio. En la segunda fase se hace predicción en sitios de la región no muestreados por medio de la técnica kriging. Kriging encierra un conjunto de métodos de predicción espacial que se fundamentan en la minimización del error cuadrático medio de predicción. Hay distintos métodos de kriging, pero en este caso se empleará el kriging ordinario.

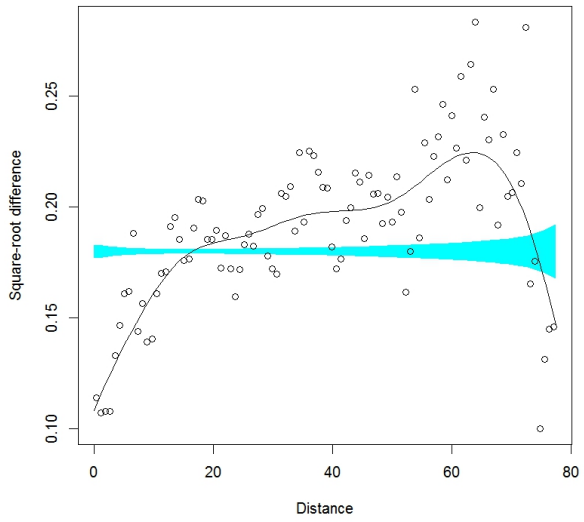

Figura 3.5: Nube de puntos del variograma del alelo q03 en escala robusta. Azul: banda de variabilidad bajo la hipótesis de independencia. Línea continua: variograma suavizado.

Supongamos que se hacen mediciones de la variable de interés Z (en este caso el alelo q03) en los puntos $\{s_j : j = 1, 2, \ldots, m\}$, de la región de estudio, es decir se tienen realizaciones de las variables $Z(s_1), \ldots, Z(s_m)$, y se desea predecir $Z(s_0)$ en el punto s_0 donde no hubo medición. En esta circunstancia, el método kriging ordinario propone que el valor de la variable puede predecirse como una combinación lineal de las n variables aleatorias de la siguiente forma:

$$\hat{Z}(s_0) = \lambda_1 Z(s_1) + \lambda_2 Z(s_2) + \lambda_3 Z(s_3) + \ldots + \lambda_m Z(s_m) = \sum_{j=1}^{m} \lambda_j Z(s_j)$$

donde los λ_j representan los pesos o ponderaciones de los valores originales. Dichos pesos se calculan en función de la distancia entre los puntos muestreados y el punto donde se va a hacer la correspondiente predicción. La suma de los pesos debe ser igual a uno para que el predictor sea insesgado.

Por tanto, este predictor es lineal,$\hat{Z}(s_0) = \lambda' \vec{Z}$, e insesgado, $E(\hat{Z}(s_0)) = E(Z(s_0))$, donde $\vec{Z} = (Z(s_1), \ldots, Z(s_m))$.

En este estudio, se obtendrán predicciones del alelo q03 en los puntos $\{s_i : i = 1, \ldots, n\}$ representados en la Figura 1.1, donde se dispone de los valores de prevalencia de EM y de las demás covariables. Se denotarán por $\hat{q}03$.

3.3. Modelos de regresión lineal con una nueva covariable

Para la elección del modelo se hace una selección de variables al modelo con las variables principales y todas las posibles interacciones entre ellas. En este caso, las variables seleccionadas para incluir en el modelo son las siguientes:

- Irradiación: I_anual
- SNP de pigmentación: q
- Longitud: $Long$
- Predicción del alelo $HLA - DRB1 * 03$: $\hat{q}03$
- Interacción de la irradiación y la longitud: $I\_anual * Long$
- Interacción de la irradiación y la predicción del alelo $HLA-DRB1*03$: $I\_anual * \hat{q}03$
- Interacción del SNP de pigmentación y la predicción del alelo $HLA - DRB1 * 03$: $q * \hat{q}03$

Se exponen en la Tabla 3.2 los resultados obtenidos para este modelo.

r^2: 0.76 / r^2 ajustado: 0.65 / AIC: 21.63			
Variables	Estimación	Error Típico	Significación
$Intercept$	8.12	4.94	0.12
I_anual	-5.13e-03	1.84e-03	0.01
q	10.14	10.64	0.35
$Long$	-0.18	0.06	0.01
$\hat{q}03$	-107.10	40.49	0.02
$I\_anual * Long$	5.10e-05	2.02e-05	0.02
$I\_anual * \hat{q}03$	0.03	0.01	0.03
$q * \hat{q}03$	-0.01	89.15	0.25

Tabla 3.2: Valores de la estimación de los coeficientes de modelo, del error típico de la estimación y de la significación de los coeficientes, considerando una nueva covariable.

En los resultados de la Tabla 3.2, la pendiente negativa de la irradiación y la longitud encajan con lo publicado y lo estudiado anteriormente, esto es, a mayor irradiación o mayor longitud, menor prevalencia de EM.

La pendiente negativa de $\hat{q}03$ no se corresponde con lo esperado. Esta variable $\hat{q}03$ es la frecuencia de un alelo de riesgo de EM y a mayor frequencia nos esperaríamos mayor prevalencia, tal y como se ve en la bibliografía y en nuestras correlaciones iniciales ($r = 0.49$, $p - valor = 0.01$). Esta variable está correlacionada con la longitud ($r = -0.72, p - valor < 0.01$), lo que podría implicar un problema de colinealidad. Sin embargo, se han calculado los factores de inflación de la varianza y estos no llegan a superar el "límite de riesgo".

La interacción positiva entre irradiación y $\hat{q}03$ implica que la pendiente en principio negativa de $\hat{q}03$ se vuelve positiva en valores medio-altos de irradiación (a partir de 2840). Este umbral lo superan la mitad sur de Europa. Recíprocamente, la pendiente negativa de la irradiación se convierte en positiva en valores altos de $\hat{q}03$ (mayores que 0.135). Este umbral lo supera Italia, Irlanda y Reino Unido. Debido a que, nuevamente, la población de Cerdeña muestra la frecuencia más alta del alelo de riesgo $HLA - DRB1 * 03$, la interpretación de esta interacción puede estar influenciada por este outiler.

La otra interacción significativa (irradiación por longitud) ilustra bien lo que parece verse en el mapa (Figura 3.6), el gradiente SE-NW. La influencia de la irradiación sobre la prevalencia es negativa en todo el rango de longitud

analizado si bien es más fuerte en valores bajos de longitud y se vuelve más suave en valores altos (más al este). La interacción también implicaría que en valores altos de radiación (a partir de 3680) la pendiente de la longitud se vuelve positiva (mayor prevalencia al este, a mayor longitud), pero de nuevo, puede explicarse por el efecto de Cerdeña.

En la Figura 3.6 se representa la superficie de predicción de la EM atendiendo al modelo de la Tabla 3.2. Se refleja un gradiente SE-NW, aunque se detecta un agudo valor de prevalencia en Cerdeña que interfiere en las predicciones del modelo.

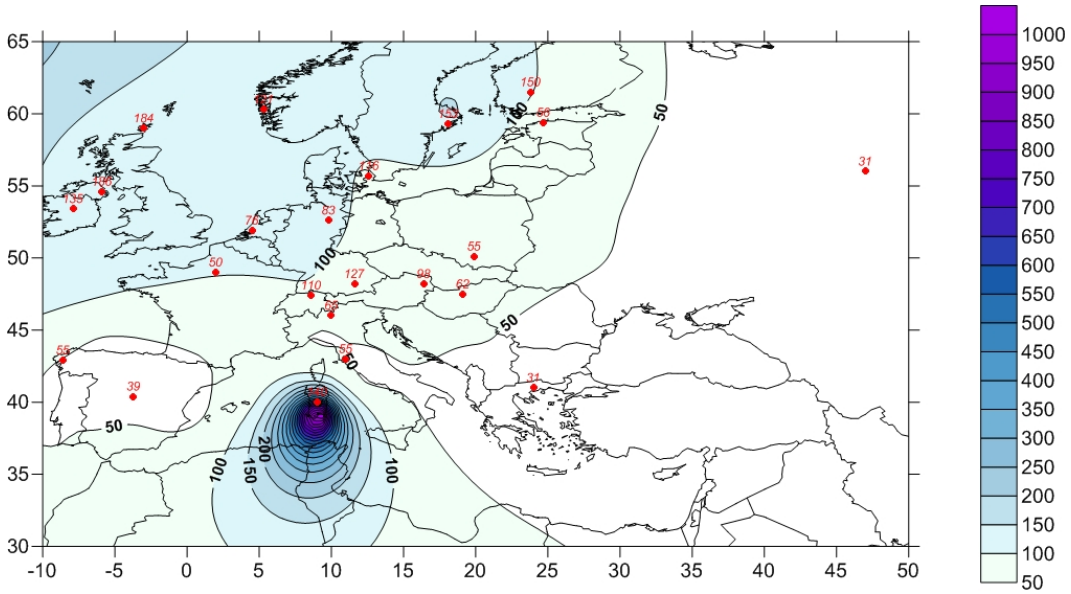

Figura 3.6: Superficie de tendencia de la prevalencia de EM con el modelo estimado incluyendo como variable explicativa el alelo q03.

Se ha hecho la diagnosis y la validación del modelo y han resultado correctos, aunque, de nuevo, la observación de Cerdeña influye en la interpretación de los resultados. Se analizarán a continuación los datos eliminando la observación de Cerdeña.

Modelo de regresión lineal excluyendo la observación de Cerdeña

Para elegir el modelo apropiado, se toma el modelo con las variables principales y todas las posibles interacciones entre ellas, y se hace una selección de variables con la función *step*. Las variables seleccionadas para el modelo son las siguientes:

- Irradiación: I_anual
- SNP de pigmentación: q
- Longitud: $Long$
- Predicción del alelo HLA-DRB1*03: $\hat{q}03$
- Interacción de la irradiación y la predicción del alelo HLA-DRB1*03: $I\_anual * \hat{q}03$
- Interacción del SNP de pigmentación y la predicción del alelo HLA-DRB1*03: $q * \hat{q}03$

En la siguiente matriz se muestran las correlaciones que presenta las variables del modelo elegido:

$$B = \begin{pmatrix} & lp & I\_anual & Long & q & \hat{q}03 \\ lp & 1.00 & -0.47 & -0.38 & -0.16 & 0.49 \\ I\_anual & -0.47 & 1.00 & -0.12 & 0.76 & 0.13 \\ Long & -0.38 & -0.12 & 1.00 & -0.17 & -0.72 \\ q & -0.16 & 0.76 & -0.17 & 1.00 & 0.29 \\ \hat{q}03 & 0.49 & 0.13 & -0.72 & 0.29 & 1.00 \end{pmatrix}$$

Los resultados obtenidos se presentan en la Tabla 3.3.

r^2: 0.79 / r^2 ajustado: 0.70 / AIC: 16.93			
Variables	Estimación	Error Típico	Significación
$Intercept$	1.08	2.67	0.69
I_anual	-2.79e-03	1.09e-03	0.02
q	21.75	11.76	0.08
$Long$	-0.04	0.01	0.004
$\hat{q}03$	-56.44	23.81	0.03
$I\_anual * \hat{q}03$	0.02	0.01	0.05
$q * \hat{q}03$	-247.20	11.00	0.04

Tabla 3.3: Valores de la estimación de los coeficientes de modelo, del error típico de la estimación y de la significación de los coeficientes del modelo sin considerar Cerdeña.

Analizando los resultados que se muestran en la Tabla 3.3, se puede ver que los efectos de irradiación y longitud se mantienen. La pigmentación (q) aparece con pendiente positiva, cercana a la significación, pero debemos tener en cuenta que su relación con la prevalencia depende también de los valores del SNP de riesgo (interacción $q * \hat{q}03$).

El alelo de riesgo $HLA - DRB1 * 03$ ($\hat{q}03$) muestra una pendiente negativa en principio opuesta a lo esperado para esta variable. Sin embargo la interpretación del papel de este alelo implica también su participación en dos interacciones.

La interacción de irradiación y $\hat{q}03$ presenta un pendiente positiva. El efecto (negativo) de la irradiación es más acusado en valores bajos del alelo $HLA-DRB1*03$. Esto podría apuntar a que el papel de la radiación sobre la síntesis de la VD sea más importante en zonas de bajo riesgo genético de EM. La VD podría ser un factor ambiental de riesgo adicional pero insignificante frente a ciertos alelos de riesgo genético.

Recíprocamente, a partir de valores medios de irradiación (umbral 2621) la pendiente de $\hat{q}03$ se vuelve positiva. Por debajo de ese umbral solo hay siete poblaciones, justo las del Norte, en todo el resto la pendiente sería positiva, encajando con lo que se esperaría de este alelo. Es decir, según nuestros datos el alelo de riesgo de EM ayudaría a explicar la prevalencia observada de EM pero especialmente en las zonas en las que la prevalencia no es tan extrema, contrastando en parte con el trabajo de Handel et al. [6] que apuntaba a este alelo como el más claramente relacionado con el gradiente latitudinal en

prevalencia. En la alta prevalencia observada en los países del Norte sin duda habría que tener en cuenta la frecuencia de otros alelos de riesgo no incluidos en este estudio.

Se muestra una pendiente negativa para la interacción de q y $\hat{q}03$. La pendiente positiva del SNP de pigmentación se encuentra sólo en las poblaciones con una frecuencia para el alelo de riesgo menor que 0.09: Suecia, Estonia, Finlandia, Rusia y Grecia. Son el eje más oriental del estudio. Recordemos que $\hat{q}03$ está negativamente correlacionada con la longitud ($r = -0.72$) así que apuntaría a un posible papel de la pigmentación en el este de Europa o zonas de baja frecuencia de $\hat{q}03$ (bajo riesgo). Nuevamente volvemos a apuntar a un posible papel de la VD (vía capacidad de síntesis en función de la pigmentación) solo en casos o zonas de bajo riesgo genético.

Se ilustra en la Figura 3.7 la superficie de tendencia de las predicciones de la prevalencia de EM para el modelo correspondiente a la Tabla 3.3. Excluyendo la observación de Cerdeña, las predicciones del modelo se comportan conforme los datos de prevalencia iniciales.

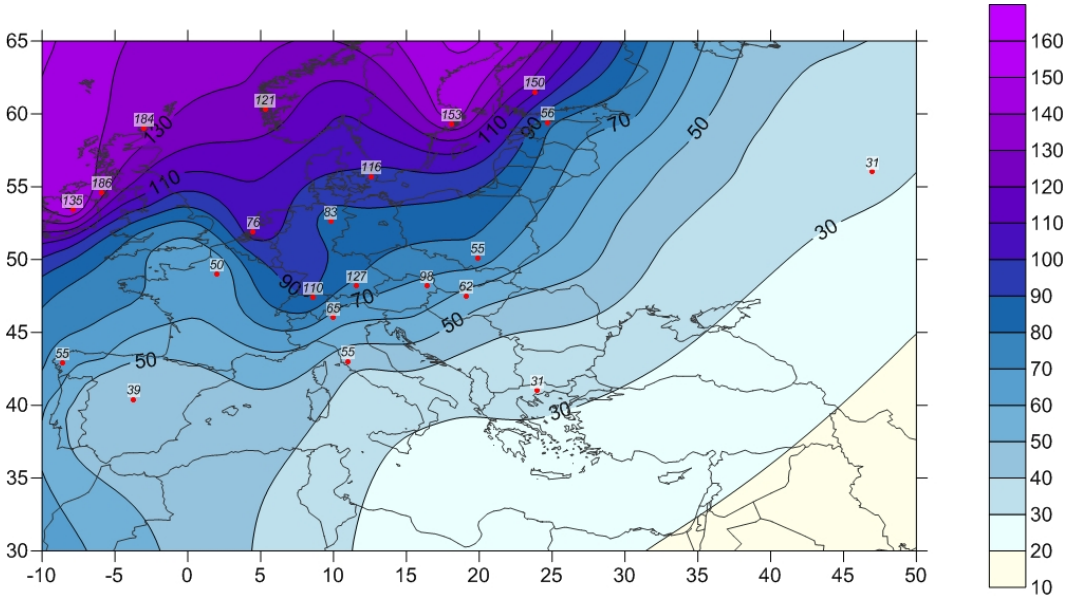

Figura 3.7: Superficie de tendencia de la prevalencia de EM con el modelo estimado.

Capítulo 4

Conclusiones

Las técnicas de regresión combinadas con los métodos de la estadística espacial han resultado útiles, en el contexto del estudio abordado en este trabajo, para caracterizar la distribución espacial de una enfermedad compleja, como la esclerosis múltiple.

Pese a las limitaciones encontradas tanto desde el punto de vista práctico como desde le punto de vista genético, resultaría de interés un estudio más profundo de la metodología utilizada en este trabajo dado que su potencialidad en este ámbito es evidente. En concreto, la inclusión de técnicas no paramétricas en el análisis de las tendencias espaciales y la exploración de otros posibles métodos de corrección del mal alineamiento de los datos (covariables observadas en distintos puntos), así como su impacto en las estimaciones obtenidas, permitirían caracterizar la distribución espacial de la enfermedad considerada y realizar predicciones en zonas donde no se tienen observaciones, teniendo en cuenta la variabilidad a la que están sujetas las estimaciones y predicciones obtenidas.

En relación a las limitaciones prácticas del estudio, se debe resaltar que las conclusiones derivadas han de ser analizadas con precaución, ya que los resultados mostrados corresponden a un mero estudio preliminar del posible papel de varios factores genéticos sobre la distribución de la prevalencia de la EM en Europa. La inclusión de datos genéticos ha limitado notablemente el número de observaciones disponibles, especialmente si se compara, por ejemplo, con un estudio en el que solo se hubiera valorado la relación de la prevalencia con la irradiación y las coordenadas geográficas.

Desde el punto de vista genético las limitaciones van más allá del bajo número de observaciones. El posible papel de la pigmentación y su inter-

acción con la irradiación se ha valorado indirectamente y de un modo muy simplificado, a través de un único SNP de un gen relacionado con la pigmentación. Incluir información más detallada de otros genes relacionados con la pigmentación o incluir información fenotípica acerca de este rasgo permitiría valorar la hipótesis inicial con mayor exactitud.

Además, teniendo en cuenta que estamos asumiendo que el papel de la irradiación y la pigmentación sobre la EM es a través de la síntesis de VD, se debe destacar que no se han analizado genes que están implicados en esta ruta directamente, de los que no conocemos su distribución espacial. Por último, sólo se ha valorado el papel de un único alelo del gen HLA-DRB1, uno de los más fuertemente asociados con la enfermedad. Hay más genes implicados y sobre todo, hay otros alelos de ese mismo gen (unos de riesgo, otros protectores) que podrían ser considerados para una caracterización más precisa. En este estudio, teniendo en cuenta su carácter exploratorio y el bajo número de observaciones se ha seleccionado HLA-DRB1*03 por ser el que presentaba más evidencias de una distribución heterogénea que podía contribuir a caracterizar el gradiente en prevalencia.

Considerando todas estas limitaciones podemos apuntar las siguientes conclusiones:

- La pigmentación, si bien es mantenida en los modelo tras la selección de variables, tiene un efecto bastante sutil sobre la prevalencia, comparado con el efecto constatado para irradiación o longitud. Esto puede ser debido a las limitaciones del estudio anteriormente señaladas, especialmente a la falta de variabilidad entre poblaciones en latitudes semejantes, o a que el efecto de la pigmentación no sea significativo. Es decir, puede que el efecto de la irradiación sobre la síntesis de VD sea tan grande que el sutil efecto de la pigmentación apenas contribuya al riesgo de padecer EM en las poblaciones del norte con una irradiación extremadamente baja.

- Se confirma el papel de la irradiación y de la longitud. Estos efectos y las interacciones encontradas ilustran un gradiente SE-NW, más que el gradiente latitudinal.

- La población de Cerdeña debe ser excluida de este tipo de aproximaciones. Su valor extremo de prevalencia (muy estudiado y con posible base genética), en el extremo sur de Europa, tiene un peso excesivo en

los modelos de regresión e interfiere en la interpretación correcta de los coeficientes.

- Alguna de las interacciones encontradas (o sugeridas en algunos casos), si bien necesitarían ser confirmadas en posteriores estudios, podrían apuntar a que el papel de la irradiación (considerada siempre como *proxy* de la VD) sobre la prevalencia de la enfermedad variara a lo largo de Europa o en función de otros factores genéticos. Confirmar este punto es de especial interés en relación con el polémico tema de la utilidad de los suplementos en esta vitamina en el tratamiento de la enfermedad.

Capítulo 5

Anexo

Funciones y paquetes de R utilizadas:

- AIC
  Función para calcular Criterio de Información de Akaike (2.2) para uno o varios objetos del modelo ajustado.

- cooks.distance
  Función que calcula la distancia de Cook para analizar los datos influyentes en el modelo.

- step
  La función step (2.1.3) selecciona un modelo de regresión a partir del criterio de información de Akaike. Busca el modelo que tenga el mínimo AIC eliminando aquellas variables que hacen éste no mejore.

- sm.variogram
  Función del paquete sm que permite realizar el contraste de independencia en datos espaciales utilizando el test de Diblasi y Bowman (2001). El código de esta función ha sido modificado para utilizar distancias en coordenadas geométricas.

- geoR
  Este paquete provee funciones para el análisis de datos geoestadísticos usando el software R. De este paquete se han utilidazado diversas funciones: sm.variogram, as.geodata, expand.grid, variog, variofit y krige.conv.

Bibliografía

[1] Acheson E.D., Bachrach C.A. y Wright F.M.(1960) Some comments on the relationship of the distribution of multiple sclerosis to latitude, solar radiation, and other variables. *Acta Psychiatr Scand*, 147, 132-147.

[2] Diblas, A. y Bowman, A.W. (2001) On the use of the variogram in checking for independence in spatial data. *Biometrics*, 57, 211-218.

[3] Diggle,P.J. y Ribeiro,P.J.(2007) *Model-based Geostatistics*. Springer, New York.

[4] Faraway, J.J.(2004) *Linear models with R*. Chapman and Hall, London.

[5] Freedman D.M., Dosemeci M. y Alavanja M.C.(2000) Mortality from multiple sclerosis and exposure to residential and occupational solar radiation: a case-control study based on death certificates. *Occupational and Environmental Medicina*, 57, 418-421.

[6] Handel A.E., Handunnetthi L., Giovannoni G., Ebers G.C. y Ramagopalan S.V.(2010) Genetic and environmental factors and the distribution of multiple sclerosis in Europe. *European Journal of Neurology*, 17, 1210-1214.

[7] Jablonski N.G. y Chaplin G.(2000) The evolution of human skin coloration.*Journal of Human Evolution*, 39, 57-106.

[8] Leibowitz U., Sharon D. y Alter M.(1967) Geographical considerations in multiple sclerosis. *Brain*, 90, 871-886.

[9] Pugliatti M., Rosati G., Carton H., Riise T., Drulovic J., Vécsei L. y Milanov I.(2006) The epidemiology of multiple sclerosis in Europe. *European Journal of Neurology*, 13, 700-722.

[10] Sanchez Sellero,C.A. (2010) *Material de Métodos de regresión.*«http://www.usc.es/campusvirtual»

[11] Schabenberger,O. y Gotway,C.A. (2005) *Statistical Methods for Spatial Data Analysis*. Chapman and Hall, Boca Ratón.

[12] Steve Simpson Jr, Leigh Blizzard, Petr Otahal, Ingrid Van der Mei y Bruce Taylor (2011) Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis. «http://jnnp.bmj.com».

[13] Sutherland J.M., Tyrer J.H. y Eadie M.J.(1962) The prevalence of multiple sclerosis in Australia. *Brain*, 85, 146-164.

[14] Swank R.L., Lerstad O., Strom A. y Backer J.(1952) Multiple sclerosis in rural Norway. Its geographic and occupational incidence in relation to nutrition. *The New England Journal of Medicine*, 246, 721-728.

[15] Van der Mie I.A.F., Ponsonby A.L., Blizzard L., Dwyer T., Taylor B.V., Kilpatrick T., Butzkueven H. y McMichael A.J.(2007) Vitamin D levels in people with multiple sclerosis and community controls in Tasmania, Australia. *Europan Journal of Neurology*, 254,581-590.

[16] Van der Mie I.A.F., Ponsonby A.L., Blizzard L. y Dwyer T.(2001) Regional variation in multiple sclerosis prevalence in Australia and its association with ambient ultraviolet radiation. *Neuroepidemiology*, 20, 168-174.