Estudio de algunas propiedades estadísticas de un diseño caso control pareado que controla el efecto confusor de la edad de aparición de la enfermedad

ARYCE LILIBETH PARDO CAMACHO

Tesis para optar al título de

Magister en Ciencias-Estadística

Director

FRANCISCO JAVIER DÍAZ CEBALLOS, Ph.D

Profesor asociado

Departamento de Estadística

Universidad Nacional de Colombia, Sede Medellín

Codirector

JUAN MANUEL ANAYA CABRERA,M.D.

Profesor titular

Escuela de Medicina

Universidad del Rosario

Universidad Nacional de Colombia, Sede Medellín

Facultad de Ciencias - Escuela de Estadística

29 de julio de 2009

Resumen

En esta tesis se estudian algunas propiedades estadísticas de un diseño caso control pareado que controla el efecto confusor de la edad de aparición de la enfermedad, (la cual se denomina edad índice y consiste en la edad de aparición de la enfermedad en el caso) cuando se utilizan las metodologías de la regresión logística condicional y el modelo de regresión de Cox estratificado, para establecer cual de estas puede brindar mejores resultados en el estudio de enfermedades genéticas complejas como lo son el MAS (Múltiples Enfermedades Autoinmunes) y SEMIMAS (Definida como el tener dos enfermedades autoinmunes), en las cuales el defecto bioquímico es causado por la alteración de múltiples genes. Finalmente, se examina la viabilidad de las metodologías propuestas para este tipo de estudios indagando la afirmación propuesta por Gauderman et al (1999), acerca de la interpretación que se le da al parámetro de regresión correspondiente a una variable genética "dicotómica". Es decir que la razón de hazards es igual a e^{β} para cualquier edad fija, por lo tanto el hazards de enfermarse en esa edad para los individuos que tienen el genotipo de riesgo es e^{β} veces el hazards para los individuos que no lo tienen.

Palabras claves: Regresión Logística Condicional, Modelo de Regresión de Cox estratificado, MAS, SEMIMAS, Marcadores, Razón de riesgo.

Abstract

In this thesis has been studied some statistical properties about a matched case-control design, which controls the confusor effect of the age of the disease appearance (which is denominated as age index and based on the age of the disease appearance in the (aforementioned) case) when using the methodologies of the conditional logistic regression and the stratified Cox regression model to establish which of those previously mentioned can perform better outcomes in the study of complex genetic diseases as MAS (Multiple Autoimmune Diseases) and SEMIMAS (defined as having two autoimmune diseases), in which the biochemical defect is caused by the alteration of multiple genes. Finally, it is tested the feasibility of the methodologies proposed for this kind of studies; by looking into the statement made by Gauderman et al (1999), with regard to the interpretation that has been given to the parameter of regression related to a genetic variable "Dummy". That means the Hazards rate is equal to e^{β} for any fixed age, therefore the Harzards of getting sick at that age for the individuals who have the risk genotype is e^{β} times the Hazards for the individuals who do not have it.

Keywords: Conditional Logistic Regression, Stratified Cox Regression Model, MAS, SEMIMAS, Markers, Hazard Ratio.

Índice general

Índice de Tablas

CAPÍTULO 1

Introducción

Para una enfermedad como el cáncer con edad de aparición variable, se considera al riesgo relativo (la cual es una medida del tamaño del efecto del factor de riesgo sobre el riesgo de sufrir la enfermedad) el parámetro genético de interés para medir las tasas de incidencia para una edad específica. La razón de odds para un diseño pareado caso control, es un estimador consistente del riesgo relativo, cuando la enfermedad es de baja frecuencia en la población. Suministrando controles que son aleatoriamente seleccionados del grupo de riesgo comprendiendo aquellos miembros de la población en riesgo que están libres de enfermedad en la edad en la cual el caso fué afectado (Gauderman et al, 1999).

Un problema que se presenta cuando se estudian enfermedades de aparición variable es buscar los controles adecuados. Cuando se va hacer un estudio caso control, lo que se busca es que los controles sean lo más parecidos a los casos, pero el problema

es que cuando el control no está enfermo, no se sabe si este se encuentra aliviado de la enfermedad en estudio debido a que no se va a enfermar, o porque no ha pasado suficiente tiempo para enfermarse. En este punto es donde está el problema de la edad de aparición de la enfermedad como variable de confusión. Para resolver el problema de cómo controlar la edad de aparición de la enfermedad, Gauderman et al, propusieron buscar un control que haya estado aliviado a la edad en la que el caso se enfermó, controlando así la edad de aparición de la enfermedad, es decir controlando el efecto confusor de dicha edad. En este diseño los controles seleccionados que se emplean son hermanos (o primos) de los casos. El que se tome como control a un hermano, tiene varias ventajas puesto que se garantiza que tanto el caso como el control se encuentran expuestos a condiciones ambientales similares, y además como comparten información genética se logra controlar en gran medida el problema de la confusión causada por la variable raza. El diseño propuesto por Gauderman aplica para enfermedades de baja frecuencia, en donde la edad de aparición de la enfermedad es variable, es decir que se hace necesario realizar un estudio caso control pareado, en donde se controla la edad de aparición de la enfermedad.

Actualmente, en el área investigativa de la medicina y áreas afines se presenta un especial interés por conocer la relación entre la exposición a ciertos factores considerados en algunas ocasiones de riesgo y una determinada enfermedad. Es por esta razón que resulta útil conocer cual es la metodología más indicada para cuantificar el efecto de dichos factores. En el presente trabajo se observan los resultados obtenidos al aplicar la regresión logística condicional y el modelo de regresión de Cox estratificado, para estudios caso control basados en familias, en donde se controla la edad de aparición de la enfermedad (edad índice). Para la aplicación de estas metodologías se contó con la ayuda del Dr. Juan Manuel Anaya, quien nos permitió utilizar la información genética junto con algunos datos de carácter clínico de pacientes que padecen de MAS (Múltiples Enfermedades

Autoinmunes) recolectada por él y su grupo de colaboradores, a lo largo de 7 años. La razón por la cual se tardó tanto tiempo en recopilar la información de estos pacientes, es debido a la baja frecuencia de esta enfermedad dentro de la población, convirtiéndola por lo tanto en una muestra muy valiosa para el estudio de esta enfermedad. Adicionalmente, se planteó también el trabajar con SEMIMAS (Definida como el tener dos enfermedades autoinmunes), aprovechando que los datos disponibles así nos lo permiten. Posteriormente, se calcula el tamaño del efecto de los alelos, sobre el riesgo de sufrir Múltiples Enfermedades Autoinmunes (MAS), permitiendo evaluar la importancia clínica de dichos alelos, comprobando la siguiente afirmación "la razón de odds es una razón de hazards en un estudio caso control pareado por la edad de aparición de la enfermedad" (Gauderman et al, 1999).

El presente estudio tiene como objetivo central, dilucidar la conexión entre el modelo de regresión logística condicional y el modelo de regresión de Cox estratificado, en estudios casos control aplicados a familias en enfermedades con edad de aparición variable. Lo que permite comprobar la siguiente afirmación hecha por Gauderman "La razón de hazard es igual a e^{β} para cualquier edad fija, por lo tanto el hazard de enfermarse en esa edad para los individuos que tienen el genotipo de riesgo es e^{β} veces el hazard para los individuos que no lo tienen".

El capítulo 1 presenta una introducción al problema.

El capítulo 2 contiene algunas definiciones básicas referentes a la regresión logística condicional y el modelo de regresión de Cox estratificado, las cuales permitirán la comprensión de la notación que se va a utilizar en el desarrollo de este trabajo. Además se plantean algunos conceptos básicos de genética con el fin de ambientar la comprensión posterior de la aplicación realizada.

El capítulo 3 describe la metodología utilizada en este estudio, para la obtención de resultados analíticos, que permitieron desarrollar los objetivos propuestos.

En el capítulo 4 se muestran los resultados de la aplicación realizada a los datos suministrados por La Corporación para Investigaciones Biológicas (CIB), en pacientes que sufren de MAS (Múltiples Enfermedades Autoinmunes) y SEMIMAS (Definida como el tener dos enfermedades autoinmunes).

El capítulo 5 presenta las conclusiones del estudio, tanto para la parte analítica como para la aplicación.

En los anexos, se muestran los programas desarrollados, para este trabajo, los nombres de los marcadores genéticos con los que se contaban, algunas tablas que permiten describir la muestra de análisis para cada una de las metodologías dependiendo de la enfermedad en estudio, y finalmente el formato de la carta de consentimiento que cada una de las personas incluidas en este estudio firmo.

CAPÍTULO 2

Marco Teórico

En este capítulo, se encuentra la sustentación teórica de los métodos estadísticos implementados en cada una de las etapas del análisis. Los cuales son Regresión Logística Condicional y el Modelo de Regresión de Cox estratificado. Adicionalmente se darán a conocer algunos conceptos básicos adicionales que resultan ser de interés para este estudio. Por lo tanto, a continuación se presentará una descripción resumida de las técnicas estadísticas que se van a utilizar.

2.1. Diseño

2.1.1. Diseño de Gauderman

El diseño propuesto por Gauderman en 1999, puede ser aplicado a enfermedades de baja frecuencia, como por ejemplo el cancer o las enfermedades autoimunes, en

donde la edad de aparición de la enfermedad es variable, lo que significa que puede presentarse en cualquier etapa de la vida (niñez, adultez o vejez), para estudiarlas y comprenderlas mejor se hace necesario realizar un estudio de casos y controles pareado, en donde se controle la edad de aparición de la enfermedad, (denominada edad índice) mediante la siguiente estrategia; para la selección de los controles es necesario que estos se encuentren aliviados a la edad en la que el caso se enfermó, lo que permite controlar la edad de aparición de la enfermedad, es decir que se controla el efecto confusor de dicha edad, esto es necesario debido a que cuando el control no está enfermo, no se sabe si este se encuentra aliviado de la enfermedad en estudio debido a que no se va a enfermar, o porque no ha pasado suficiente tiempo para enfermarse. Gauderman también propone que los controles seleccionados pueden ser hermanos (o primos) de los casos, lo que suministra varias ventajas debido a que se garantiza que tanto el caso como el control se encuentran expuestos a condiciones ambientales similares, y comparten información genética, lo que permite controlar en gran medida el problema de la confusión causada por las variables habitat y raza.

2.1.2. Controles Hermanos

Al dejar de considerar a la población origen como la población entera y considerar solamente a los familiares del caso como controles potenciales, el investigador empareja a cada caso a unos o más controles del él (Gauderman, Witte y Thomas, 1999).

2.1.3. Ventajas y Desventajas de los Controles Hermanos

- La ocurrencia de la enfermedad en el caso puede hacer a sus parientes más dispuestos a participar en la investigación. Generando así, una mayor voluntad del control a completar con mas cuidado el cuestionario de factores de riesgo.
- Reduce el costo de la investigación.

Sin embargo se presenta la desventaja, que no todos los casos tendrán un hermano elegible y dispuesto.

2.1.4. Análisis Estadístico del Diseño de Gauderman

Como se tiene un estudio pareado caso control, cuya respuesta es dicotómica (tener o no la enfermedad en estudio, lo que denominaremos ser caso o control respectivamente) la herramienta más adecuada es la regresión logística condicional y las variables independientes corresponden a las variables genéticas o ambientales de interés, las cuales también son dicotómicas con el fin de poder interpretar de una manera interesante el coeficiente de regresión; el coeficiente de regresión de la variable genética permite calcular una razón de odds en la forma usual, exponenciando el coeficiente de regresión. Esta razón de odds se interpreta como una razón de hazards; el hazards de enfermarse a una edad determinada, para una persona que tiene el genotipo de riesgo, dividido por el hazard de enfermarse a la misma edad entre las personas que no tienen el genotipo de riesgo.

$$\begin{aligned} \text{logit}\left[\mathrm{P}\left(y=1|\mathrm{g}\right)\right] &= \alpha_i + \beta G\left(g\right) \\ &= \ln\left(\frac{\Pr(y=1|\text{g riesgo})}{\Pr(y=1|\text{g no riesgo})}\right) \\ &= \frac{e^{\alpha_i+\beta_1}}{e^{\alpha_i+\beta_0}} \\ &= e^{\beta} \end{aligned} \tag{2.1}$$

y Denota el estatus de la enfermedad $y = 1$ se refiere a ser un caso, mientras que $y = 0$ corresponde a un control

g Corresponde al genotipo en algún locus de interés

$G(g)$ Denota una covariable genética, la cual puede tomar los valores de 1 o de 0, dependiendo de si el alelo en el locus es clasificado como normal o mutante

β Corresponde al log del riesgo relativo de una mutación

e^{β} Es la razón de odds, que compara los individuos expuestos al factor de riesgo con los que no se encuentran expuestos al factor de riesgo

β_0 y β_1 Son los parámetros específicos de cada sujeto

En este diseño, cada caso es pareado con un familiar, en esta oportunidad con un hermano. La variable de pareamiento que se utiliza es la edad de aparición de la enfermedad. Se buscan controles que hayan estado aliviados a la edad en la que el caso se enfermo, controlando así el efecto confusor de la edad de aparición de la enfermedad.

2.2. Conceptos Básicos

2.2.1. Función de Supervivencia

La función, denotada por $S(t)$, es definida como la probabilidad de sobrevivir más allá del tiempo t, (Donde la T denota el tiempo de supervivencia) y se obtiene mediante la siguiente expresión:

$$S(t) = \mathrm{P}(\mathrm{T} > \mathrm{t}) \tag{2.2}$$

La función de supervivencia es el complemento de la función de distribución acumulada, de T:

$$S(t) = \mathrm{P}(\mathrm{T} > \mathrm{t}) = 1 - F(t) = \int_t^{\infty} f(x)\, dx \tag{2.3}$$

Entonces,

$$f(t) = -\frac{dS(t)}{dt} \tag{2.4}$$

$f(t)$, es una función no negativa con un área bajo la curva igual a uno.
Las propiedades básicas que presentan las curvas de sobrevivencia son las siguientes (Klein y Moeschberger, 2003). Son monótonas, funciones decrecientes iguales a uno y cero, que toman el valor de cero cuando se aproximan a infinito.

2.2.2. Función Hazard

Una cantidad, fundamental en el análisis de supervivencia es la función hazard, la cual se encuentra definida como sigue (Frees, 2004).

$$\begin{aligned} h(t) &= \frac{\text{Función de Densidad de Probabilidad}}{\text{Función de Supervivencia}} \\ &= \frac{-\frac{\partial}{\partial t}\mathrm{P(T>t)}}{\mathrm{P(T>t)}} \\ &= -\frac{\partial}{\partial t}\mathrm{lnP}\left(\mathrm{T}>\mathrm{t}\right) \end{aligned} \tag{2.5}$$

La anterior ecuación nos define la probabilidad instantánea de fallar, condicionado a una supervivencia en un tiempo t. La única restricción que presenta $h(t)$ es que esta es no negativa, es decir $h(t) \geq 0$

Otra función que resulta de interés es la función de hazard acumulada, la cual se encuentra definida como:

$$H(t) = \int_0^t h(s)\, ds \tag{2.6}$$

$$\begin{aligned} \int_0^t h(s)\, ds &= \int_0^t \frac{F'(s)}{1-F(s)} ds \\ &= -\ln\left(1-F(s)\right)|_0^t \\ &= -\ln\left(1-F(t)\right) + \ln\left(1-F(0)\right) \\ &= -\ln\left(1-F(t)\right) \end{aligned} \tag{2.7}$$

Esta función también se puede expresar de la siguiente forma:

$$\mathrm{P}\left(\mathrm{T}>\mathrm{t}\right) = \exp\left(-H(t)\right) \tag{2.8}$$

Definamos δ como una función indicadora para censura a la derecha, como:

$$\delta = \begin{cases} 1 \text{ si T es censurado} \\ 0 \text{ en otro caso} \end{cases} \tag{2.9}$$

La verosimilitud puede ser expresada en términos de la función hazard y de la hazard acumulada de la siguiente forma:

$$\begin{cases} \mathrm{P}\left(\mathrm{T} > \mathrm{t}\right) \text{ si T es censurado} \\ -\frac{\partial}{\partial t}\mathrm{P}\left(\mathrm{T} > \mathrm{t}\right) \text{ en otro caso} \end{cases} \tag{2.10}$$

$$\begin{aligned} &= \left(\mathrm{P}\left(\mathrm{T} > \mathrm{t}\right)\right)^{d} \left(h\left(t\right)\mathrm{P}\left(\mathrm{T} > \mathrm{t}\right)\right)^{1-d} \\ &= h\left(t\right)\exp\left(-H\left(t\right)\right) \end{aligned} \tag{2.11}$$

Aquí se asume que la función hazard puede ser escrita en términos del producto de la hazard de "línea base" y una función de una combinación de variables explicativas.

$$h\left(t\right) = h_0\left(t\right)\exp\left(X_i'\beta\right) \tag{2.12}$$

donde $h_0\left(t\right)$ es la hazard de línea base.

Esto es conocido como modelo de hazard proporcional.

Lo anterior lo podemos ver más claramente por medio del siguiente ejemplo si tomamos las funciones hazard de dos conjuntos de covariables X_1 y X_2, se obtiene:

$$\frac{h\left(t|X_1\right)}{h\left(t|X_2\right)} = \frac{h_0\left(t\right)\exp\left(X_1'\beta\right)}{h_0\left(t\right)\exp\left(X_2'\beta\right)} = \exp\left(\left(X_1 - X_2\right)'\beta\right) \tag{2.13}$$

Como se puede observar esta razón resulta ser independiente del tiempo t.
Un valor de 1 para la razón hazard corresponde a la igualdad entre ambos. El valor mínimo de la razón de hazard es cero, el valor máximo es infinito. Si la razón hazard es superior a uno el grupo situado en el numerador resulta perjudicial. Si la razón hazard es inferior a uno el grupo situado en el numerador resulta protector.
Asumiendo que cada T_i sigue un modelo de hazard proporcional, la función de verosimilitud es:

$$\begin{aligned} L\left(\beta, h_0\right) &= \prod_{i=1}^{n} h\left(T_i\right)^{1-\delta_i}\exp\left(-H\left(T\right)\right) \\ &= \prod_{i=1}^{n} h\left(T_i\right)\exp\left(X_i'^{\beta}\right)^{1-\delta_i} exp\left(-H_0\left(T_i\right)\exp\left(X_i'^{\beta}\right)\right) \end{aligned} \tag{2.14}$$

Maximizando esto en términos de h_0 tenemos la función de verosimilitud parcial:

$$L_P\left(\beta\right) = \prod_{i=1}^{n}\left(\frac{\exp\left(X_i'^{\beta}\right)}{\sum\limits_{j\in R(T_i)}\left(X_j'^{\beta}\right)}\right)^{1-\delta_i} \tag{2.15}$$

donde $R\left(t\right)$ es el conjunto de todos los $\{T_1, \cdots T_n\}$ siendo $T_i \geq t$, que en realidad son todos los sujetos que se encuentran en el estudio en el tiempo t (Frees, 2004). De la anterior ecuación se encuentra que la inferencia para los coeficientes de la regresión depende únicamente del rango de las variables $\{T_1, \cdots T_n\}$ y no de los valores actuales.

2.2.3. Estudios Caso-Control

Los estudios de casos y controles consisten en la comparación de un grupo de enfermos con uno o más grupos de controles o testigos que no sufren de la enfermedad en estudio, en relación con la frecuencia de variables o eventuales factores causales, o con la exposición previa a algunos agentes.
Este tipo de estudios son relativamente simples, de bajo costo, permiten explorar varias hipótesis simultáneamente y son el único sistema utilizable en enfermedades de baja frecuencia en las que se desee estudiar factores causales. Estos estudios presentan una gran aplicación en el campo clínico para establecer la existencia de factores de riesgo, asociados con atributos, hábitos o uso de medicamentos por parte de pacientes.(Guerrero, González, Medina, 1981).

Los estudios de casos y controles aplicados a epidemiología genética, con frecuencia se llaman simplemente "estudios de asociación", y se usan para investigar la relación entre una exposición y una enfermedad.
El muestreo en las dos poblaciones se hace introduciendo restricciones para que las muestras de ambas queden en estratos homogéneos con respecto a alguna variable de confusión, por ejemplo la edad, entonces se muestrearía de tal modo que los enfermos y los no enfermos quedaran en estratos homogéneos según grupos de edad.

2.2.4. Ventajas de los estudios caso-control

En términos generales los estudios de casos y controles son adecuados indicados:

1. Cuando la enfermedad es rara. Un ejemplo es en enfermedades tales como el cáncer, cuya incidencia es baja.

2. Cuando se quiere hacer la exploración simultánea de varios factores. Cuando no se tiene una hipótesis concreta, es preferible hacer un estudio de casos y

controles para buscar asociaciones significativas que posteriormente pueden ser verificadas con otros estudios.

3. Son más baratos, se pueden realizar en menos tiempo, y son más fáciles de ejecutar que los estudios de cohortes.

2.2.5. Desventajas de los estudios caso-control

1. Son poco útiles cuando la frecuencia de exposición al factor causal investigado es muy baja.

2. No producen estimativos directos de la incidencia de la enfermedad en los individuos expuestos y no expuestos.

3. Puesto que en ocasiones, se trata de averiguar eventos que ocurrieron en el pasado, el aspecto relacionado con la memoria o recuerdos del entrevistado adquiere una importancia capital.
   Un problema de más difícil tratamiento es aquel donde los casos, por el hecho de estar sufriendo la enfermedad, tienden a ser mejores colaboradores y recordar mejor los eventos que pudieron haberlos llevado al estado actual de salud. Se introduce entonces un sesgo por el cual los casos aparecen artificialmente con una mayor frecuencia de exposición debido a la mejor colaboración o recuerdo de la exposición.

2.2.6. Regresión Logística

El análisis de regresión logística es utilizado como herramienta de clasificación, para determinar la clase a la que pertenece un individuo a partir de un conjunto de variables explicativas. Este instrumento estadístico de análisis multivariado, resulta

útil cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas o categóricas. En este último caso, se requiere que sean transformadas en variables "dummy" (Hosmer y Lemeshow, 1989).La variable dependiente **Y** representa la ocurrencia o no de un suceso. Podemos decir que la variable dependiente **Y** toma valor 1 si ocurre el suceso, y valor 0 si no ocurre el suceso.

Nos interesa estudiar la relación entre una o más variables independientes o explicativas: $X_1, X_2, ..., X_p$ y la variable Y. El modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores $X_1 = x_1, X_2 = x_2, ..., X_p = x_p$:

$$\begin{aligned} \pi(\mathbf{x}) &= P(Y = 1 | x_1, x_2, ..., x_p) \\ &= \frac{1}{1+\exp(-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p))} \end{aligned} \tag{2.16}$$

En el modelo de regresión logística simple la probabilidad $\pi(x)$ esta dada por:

$$\pi(x) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))} \tag{2.17}$$

Cuando $(x) \to \infty$, $\pi(x) \to 0$ si $\beta_1 < 0$ y $\pi(x) \to 1$ si $\beta_1 > 0$.

2.2.7. Estimación de los parámetros utilizando el método de Máxima Verosimilitud

La estimación de los parámetros por medio del método de máxima verosimilitud, proporciona los valores de los parámetros desconocidos que maximizan la proba-

bilidad de obtener el conjunto de datos observado. El procedimiento consta de los siguientes pasos:

1. Primero se construye la función de verosimilitud. Si Y es codificada como cero o uno, entonces la expresión para $\pi(x)$ da la probabilidad condicional de que $\mathbf{Y}$ sea igual a 1 dado x y la cantidad $1-\pi(x)$ da la probabilidad condicional de que $\mathbf{Y}$ sea igual a 0 dado x. Para los pares (x_i, y_i) en los cuales $y_{\text{i}} = 1$ la contribución a la función de verosimilitud es $\pi(x_i)$ y para los pares en los que $y_{\text{i}} = 0$ la contribución a la función de verosimilitud es $1-\pi(x_i)$ Por lo tanto, la contribución del par (x_i, y_i) a la función de verosimilitud es $\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$.

   Como las observaciones se asumen independientes, la función de verosimilitud es obtenida como el producto de los n términos, es decir:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{2.18}$$

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} y_i \ln(\pi(x_i)) + (1-y_i)\ln(1-\pi(x_i)) \tag{2.19}$$

Se encuentra el valor de B que maximiza $L(\text{B})$ derivando con respecto a B y el conjunto de expresiones resultantes se iguala a cero.

Así, las ecuaciones de verosimilitud son:

$$\begin{aligned} &\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \\ &\sum_{i=1}^{n} \{x_{ij}[y_i - \pi(x_i)]\} = 0 \end{aligned} \tag{2.20}$$

2.2.8. Regresión Logística Condicional

La regresión logística condicional, es una extensión del modelo de regresión logística. Este modelo es usado para el análisis de muestras estratificadas. Ejemplos de la aplicación incluye información recolectada de diferentes sitios tales como escuelas, hospitales y clínicas donde el análisis a covariables son controladas por definiciones arbitrarias de estratificación de variables, el caso mas frecuente el estudio caso control.

La idea básica es expandir el modelo de regresión logística con la inclusión de variables estratificadas. La verosimilitud para el k-ésimo estrato es obtenida como la probabilidad condicional de los datos observados en el estrato del tamaño de muestra total y el número total de casos.

En general, la regresión logística condicional asume que cada estrato tiene un intercepto característico para el modelo de regresión logística. El término "condicional" se debe a la forma de estimación de los coeficientes de regresión y no al modelo en si, el cual es un modelo de regresión logística(Elston, Olson y Palmer, 2002). A continuación, se muestra la función de verosimilitud condicional, la cual es resultado del producto de la contribución a la función de verosimilitud condicional para el k-ésimo estrato.

Suponga que se tienen k estratos, $k = 1, 2, 3, \cdots, K$, en donde n_{k1} corresponde al número de casos, es decir, sujetos con $y = 1$ y n_{k0} es el número de controles, es decir, sujetos con $y = 0$.

La función de verosimilitud condicional para el k-ésimo estrato es obtenido como:

$$l_k(\beta) = \frac{P(\text{Datos observados})}{P\left(\begin{array}{l}\text{Todas las posibles asignaciones de } n_{k1} \text{ sujetos con } y=1, \text{ es decir casos} \\ \text{y } n_{k0} \text{ sujetos con } y=0, \text{ es decir controles, para } n_k = n_{k0} + n_{k1} \text{ sujetos}\end{array}\right)} \tag{2.21}$$

A lo que hace referencia las posibles asignaciones, es al número de combinaciones que existen al de n_k (número total de sujetos) seleccionar n_{k1} (número de casos). Sea el subíndice j, el que denota alguna de estas posibles asignaciones. Para alguna asignación, se tiene que los casos $(y=1)$ van de $1, 2, 3, \cdots, n_{k1}$, y para los controles $(y=0)$ se tienen desde $n_{k1+1}, \cdots, n_k$ sujetos.

La contribución para la función de verosimilitud para el k-ésimo estrato esta dada por:

$$l_k(\beta) = \frac{\prod_{i=1}^{n_{k1}} P(\mathrm{x_{ki}}|y_{\mathrm{ki}}=1) \prod_{i=n_{k1}+1}^{n_k} P(\mathrm{x_{ki}}|y_{\mathrm{ki}}=0)}{\sum_j \left(\left(\prod_{i_j=1}^{n_{k1}} P\left(\mathrm{x_{ki_j}}|y_{\mathrm{ki_j}}=1\right)\right)\left(\prod_{i_j=n_{k1}+1}^{n_k} P\left(\mathrm{x_{ki_j}}|y_{\mathrm{ki_j}}=0\right)\right)\right)} \tag{2.22}$$

Teniendo en cuenta que la probabilidad de exposición al factor de riesgo para el caso, se encuentra dada por:

$$P(\mathrm{x_{ki}}|y_{\mathrm{ki}}=1) = \frac{P(y_{\mathrm{ki}}=1|\mathrm{x_{ki}})\, P(\mathrm{x_{ki}})}{P(y_{\mathrm{ki}}=1)} \tag{2.23}$$

La probabilidad de exposición al factor de riesgo para el control es:

$$P\left(\mathrm{x}_{\mathrm{ki}_j}|y_{\mathrm{ki}}=0\right)=\frac{P\left(y_{\mathrm{ki}}=0|\mathrm{x}_{\mathrm{ki}_j}\right)P\left(\mathrm{x}_{\mathrm{ki}_j}\right)}{P\left(y_{\mathrm{ki}}=0\right)} \tag{2.24}$$

Lo anterior se tiene al aplicar el teorema de Bayes.

$$\begin{aligned} &P\left(y_{\mathrm{ki}}=0|\mathrm{x}_{\mathrm{ki}_j}\right)=1-P\left(y_{\mathrm{ki}}=1|\mathrm{x}_{\mathrm{ki}}\right)\\ &=\frac{1}{1+\exp\left(\beta_0+\prod_{i=1}^{n_{k1}}\beta\mathrm{x}_{\mathrm{ki}_j}\right)} \end{aligned} \tag{2.25}$$

$$P\left(y_{\mathrm{ki}}=1|\mathrm{x}_{\mathrm{ki}}\right)=\frac{\exp\left(\beta_0+\prod_{i=1}^{n_{k1}}\beta\mathrm{x}_{\mathrm{ki}}\right)}{1+\exp\left(\beta_0+\prod_{i=1}^{n_{k1}}\beta\mathrm{x}_{\mathrm{ki}}\right)} \tag{2.26}$$

Teniendo en cuenta que el modelo de regresión logística, puede ser expresado como:

$$\pi\left(\mathrm{x}_{\mathrm{ki}}\right)=P\left(y_{\mathrm{ki}}=1|\mathrm{x}_{\mathrm{ki}}\right) \tag{2.27}$$

Al reemplazar se tiene:

$$l_k\left(\beta\right)=\frac{\prod_{i=1}^{n_{k1}}\pi\left(\mathrm{x}_{\mathrm{ki}}\right)\prod_{i=n_{k1}+1}^{n_k}\left(1-\pi\left(\mathrm{x}_{\mathrm{ki}}\right)\right)}{\sum_j\left(\left(\prod_{i_j=1}^{n_{k1}}\pi\left(\mathrm{x}_{\mathrm{ki}}\right)\right)\left(\prod_{i_j=n_{k1}+1}^{n_k}\left(1-\pi\left(\mathrm{x}_{\mathrm{ki}}\right)\right)\right)\right)} \tag{2.28}$$

La función de verosimilitud es el producto de $l_k(\beta)$ sobre los K estratos:

$$L(\beta) = \prod_{k=1}^{K} l_k(\beta) \tag{2.29}$$

La función de verosimilitud es entonces:

$$L(\beta) = \prod_{k=1}^{K} \left(\frac{\prod_{i=1}^{n_{k1}} \exp(\beta' \mathbf{x}_{ki})}{\sum_{j} \left(\prod_{i_j=1}^{n_{k1}} \exp(\beta' \mathbf{x}_{ki_j}) \right)} \right) \tag{2.30}$$

K Es el número total de estratos

n_{k1} Corresponde al número de casos, es decir, sujetos con $y = 1$

β' Vector de parámetros desconocidos

$\mathbf{x}_{ki}$ Vector de p variables explicativas

i_j Corresponde a la j-ésima posible asignación

2.2.9. Modelo de Regresión Cox

El análisis de supervivencia nos permite estudiar y construir modelos para analizar el tiempo que un suceso tarda en ocurrir, en los que diferentes variables de pronóstico permiten estimar el tiempo de aparición del suceso. Entre los diferentes tipos de modelos que se pueden emplear, uno de los más extendidos en medicina es el modelo de riesgos proporcionales, también conocido como modelo de Cox.

Un modelo de supervivencia es una fórmula matemática que nos permite cuantificar la probabilidad de supervivencia, dados unos determinados valores de los factores de pronóstico en un momento inicial. A partir de ese cálculo podemos estimar una probabilidad de supervivencia para un tiempo determinado (por ejemplo a 3 años) para los pacientes con unas determinadas características. Es posible también calcular riesgos relativos entre dos grupos de pacientes con diferentes valores de las variables de pronóstico. Otra alternativa que nos permite la utilización del modelo es ordenar los pacientes de peor a mejor pronóstico de acuerdo con la supervivencia estimada, o clasificarlos en diferentes grupos de pronóstico, siendo la clasificación más sencilla la que contempla dos grupos: mal o buen pronóstico.

La utilización de modelos de supervivencia para ordenar a los pacientes puede ser de gran importancia para ayudar a la toma de decisiones. Así por ejemplo, a la hora de asignar un hígado a un paciente de la lista de espera de trasplantes, la utilización de un índice conocido como índice de MELD, que no es más que un modelo de supervivencia de Cox, el cual permite ordenar los pacientes en función de la supervivencia esperada, de tal manera que si se asigna el hígado al paciente con peor pronóstico según esa ordenación, se ha empleado para tomar la decisión una valoración objetiva, repetible, e independiente de quién toma la decisión, ya que se basa únicamente en datos del paciente. El modelo de regresión de Cox es:

$$h\left(t|\mathbf{z}\right) = h_0\left(t\right)\exp(\beta'\mathbf{z}) \tag{2.31}$$

h_0 Función de riesgo de referencia

β Vector de parámetros desconocidos

z Vector de variables explicativas

A continuación se presentan algunos comentarios respecto al modelo de regresión de Cox.

$h_0(t)$ Es una función que depende del tiempo.

$e^{\beta' \mathbf{z}}$ Depende de las variables pronóstico o covariantes. Además se le conoce como hazard ratio, donde el cociente de riesgo (hazard ratio), se puede considerar equivalente al concepto de riesgo relativo.

$\beta' \mathbf{z}$ Es llamado el índice de riesgo. Cuanto mayor sea el índice de riesgo, peor supervivencia o peor pronóstico para ese perfil de valores de x.

2.2.10. Modelo de Regresión de Cox Estratificado

El modelo de regresión de Cox estratificado, es una modificación del modelo de Cox de hazard proporcionales, modelo que permite tomar controles por "estratificación" de un predictor que no satisface el supuesto de hazards proporcionales.
Supongamos que tenemos k variables que no satisfacen el supuesto de hazard proporcionales. A las cuales denotaremos $Z_1, Z_2, \ldots, Z_K$; las variables que satisfacen el supuesto de hazard proporcionales las denotaremos $X_1, X_2, \ldots, X_p$.
Para cumplir el procedimiento de Cox estratificado, es necesario definir una nueva variable, la cual llamaremos Z^*, para los $Z's$ a ser usados para la estratificación. En general, la estratificación de la variable Z^* puede tener K categorías, donde K es el número total de combinaciones (o estratos) formados después de la categorización de cada Z_i.

La forma general del modelo de regresión de Cox estratificado es:

$$h_k(t, \mathbf{X}) = h_{0k}(t) \exp\left[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right] \tag{2.32}$$

$$k = 1, 2, \ldots, K$$

Esta fórmula contiene el subíndice k que indica el k-ésimo estrato. Los estratos son definidos como las diferentes categorías de la variable de estratificación Z^*, y el número de estratos es igual a K.
La hazard de línea base o referencia $h_{0k}(t)$, puede ser diferente para cada estrato, sin embargo los coeficientes de $\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ son los mismos para cada estrato (Kleinbaum y David, 1997).

2.3. Conceptos Básicos de Genética

2.3.1. Marcador Genético

Un marcador genético es un segmento de ADN con una ubicación física identificable en un cromosoma y cuya herencia se puede rastrear. Un marcador puede ser un gen, o puede ser alguna sección del ADN sin función conocida. Dado que los segmentos del ADN que se encuentran contiguos en un cromosoma tienden a heredarse juntos, los marcadores se utilizan a menudo como formas indirectas de rastrear el patrón hereditario de un gen que todavía no ha sido identificado, pero cuya ubicación aproximada se conoce. Los marcadores se usan para el mapeo genético como el primer paso para encontrar la posición e identidad de un gen.

2.3.2. Alelo

Un alelo (del griego: allélon: uno a otro, unos a otras) es cada una de las formas alternativas que puede tener un gen, que se diferencian en su secuencia, y que se puede manifestar en modificaciones concretas de la función de ese gen. Al ser la mayoría de los mamíferos diploides estos poseen dos alelos de cada gen, uno de ellos procedente del padre y el otro de la madre. Cada par de alelos se ubica en igual locus o lugar del cromosoma.

2.3.3. Locus

Un locus (del latín locus, lugar; plural loci) es una posición fija sobre un cromosoma, como la posición de un gen o de un biomarcador (marcador genético). Una variante de la secuencia de ADN en un determinado locus se llama alelo. La lista ordenada de locus conocidos para un genoma particular se denomina mapa genético, mientras que se denomina cartografía genética al proceso de determinación del locus de un determinado carácter biológico.

2.3.4. Microsatélite

Los microsatélites son secuencias cortas de ADN, usualmente de uno a seis nucleótidos, que son repetidos en multiples tiempos. Los microsatélites son importantes para la búsqueda de marcadores genéticos (Duncan, 2004).

2.3.5. Homocigótico y Heterocigótico

Cuando ambos alelos del par de cromosomas son iguales, el individuo es homocigótico, en caso contrario, es decir, cuando ambos alelos del par de cromosomas

son diferentes, el individuo es heterocigótico.

2.3.6. Genoma Humano

El genoma humano consiste en 46 moléculas del ADN de doble cadena, cada molécula tiene un promedio de 130 millones de pares de bases organizadas linealmente entre dos columnas de azúcar-fosfato y cada una es enrollada alrededor de proteínas para formar un cromosoma (Mas, 2004).

CAPÍTULO 3

Conexión entre la Regresión logística Condicional y el Modelo de Regresión de Cox Estratificado

En este capítulo se presenta la conexión existente entre la regresión logística condicional y el modelo de regresión de Cox estratificado, utilizando la información consignada en el marco teórico, cuando se tiene un estudio caso control paraeado uno a uno, lo que significa que por cada caso se tome un control.

Primero es necesario definir algunas variaciones en la nomenclatura:

En un estudio caso control pareado $1:1$, se tiene que un individuo corresponde a la pareja caso control; en donde se tiene que los controles son denotados por $t = 1$, mientras que los casos lo son por $t = 2$.

$$y_{it} = \begin{cases} 1 \quad \text{Si el sujeto} \quad \text{t} \quad \text{de la pareja} \quad \text{i,} \quad \text{esta expuesto al factor de riesgo} \\ 0 \quad \text{Si el sujeto} \quad \text{t} \quad \text{de la pareja} \quad \text{i,} \quad \text{no esta expuesto al factor de riesgo} \end{cases} \tag{3.1}$$

Aquellas parejas caso control que cumplan que $y_{i1} + y_{i2} = 0$ ó $y_{i1} + y_{i2} = 2$, son llamadas parejas "concordantes" (en el estatus del factor de riesgo), por otra parte las parejas caso control que cumplan que $y_{i1} + y_{i2} = 1$ son llamadas parejas discordantes.

La variable independiente esta definida como:

$$\text{x} = \begin{cases} 1 \quad \text{Si el sujeto es un caso} \\ 0 \quad \text{Si el sujeto es un caso} \end{cases} \tag{3.2}$$

El modelo de regresión logística condicional para un estudio caso control pareado uno a uno es:

$$\text{logit}\left(\text{p}_{\text{it}}\right) = \alpha_i + x_{it}\beta \tag{3.3}$$

con $i = 1, 2, \cdots, n$ (Correspondiente a la pareja caso control), $t = 1, 2$ (Estado del sujeto dentro de la pareja, caso o control) α_i es un parámetro específico del sujeto, el cual se interpreta como el intercepto característico del sujeto i.

Sea $p\left(\vec{Y_i}, \vec{\beta} | T\left(\vec{Y_i}\right)\right)$ la función de masa de probabilidad de la distribución condicional de $\vec{Y_i}$ dado $T\left(\vec{Y_i}\right) = \sum_{t=1}^{T_i} y_{it}$.

La función de verosimilitud condicional considerada como función de $\vec{\beta}$, se define como:

$$l_c = \prod_{i=1}^{n} p\left(\vec{Y}_i, \vec{\beta} | T\left(\vec{Y}_i\right)\right) \tag{3.4}$$

Consideremos el caso en que $T_i = 2$, calculemos la distribución condicional de $\vec{Y}_i$ dado $T\left(\vec{Y}_i\right)$.
$T\left(\vec{Y}_i\right) = y_{i1} + y_{i2}$, en donde $\vec{Y}_i = (y_{i1}, y_{i2})'$.

$$P\left(\vec{Y}_i = \vec{y}_i | T\left(\vec{Y}_i\right) = 0\right) = \begin{cases} 1 & Si \quad \vec{y}_i = (0,0)' \\ 0 & Si \quad \vec{y}_i \neq (0,0)' \end{cases} \tag{3.5}$$

$$P\left(\vec{Y}_i = \vec{y}_i | T\left(\vec{Y}_i\right) = 2\right) = \begin{cases} 1 & Si \quad \vec{y}_i = (1,1)' \\ 0 & Si \quad \vec{y}_i \neq (1,1)' \end{cases} \tag{3.6}$$

$$\begin{aligned} P\left(T\left(\vec{Y}_i\right) = 1\right) &= P\left(y_{i1} + y_{i2} = 1\right) \\ &= \sum_{j=0}^{1} P\left(y_{i1} = j\right) P\left(y_{i2} = 1 - j\right) \\ &= P\left(y_{i1} = 0\right) P\left(y_{i2} = 1\right) + P\left(y_{i1} = 1\right) P\left(y_{i2} = 0\right) \end{aligned} \tag{3.7}$$

Lo anterior se tiene usando la formula de la convolución, la cual dice que si se tienen dos variables aleatorias discretas e independientes, X e Y, si X sólo toma valores α_i $j = 0, 1, 2, \cdots$, entonces: $P\left(X + Y = \gamma\right) = \sum_j P\left(X = \alpha_j\right) P\left(Y = \gamma - \alpha_j\right)$

Continuando con el desarrollo de la ecuación (3,7) se tiene:

$$
\begin{aligned}
&= \left(1-\pi\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)\right)\pi\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)+\pi\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)\left(1-\pi\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)\right)\\
&= \left(1-\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)+\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(1-\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)\\
&= \left(\frac{1+e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}-1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)+\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1+e^{-(\alpha_i+\vec{x}'_{i2}\vec{\beta})}-1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)\\
&= \frac{e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}}{\left(\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)+\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)\right)}+\frac{e^{-(\alpha_i+\vec{x}'_{i2}\vec{\beta})}}{\left(\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)+\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)\right)}\\
&= \frac{e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}+e^{-(\alpha_i+\vec{x}'_{i2}\vec{\beta})}}{\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)+\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)}
\end{aligned}
\tag{3.8}
$$

$$
P\left(\vec{Y_i}=\vec{y_i}|T\left(\vec{Y_i}\right)=1\right)=0 \quad si \quad \vec{y_i}\neq(0,1)' \quad y \quad \vec{y_i}\neq(1,0)' \tag{3.9}
$$

$$
P\left(\vec{Y_i}=(0,1)'\,|T\left(\vec{Y_i}\right)=1\right)=P\left(y_{i1}=0,y_{i2}=1|T\left(\vec{y_i}\right)=1\right) \tag{3.10}
$$

$$
\begin{aligned}
&= \frac{P(y_{i1}=0,y_{i2}=1\ ,\ y_{i1}+y_{i2}=1\ )}{P(T(\vec{y_i})=1)}\\
&= \frac{P(y_{i1}=0,y_{i2}=1\ \ )}{P(T(\vec{y_i})=1)}\\
&= \frac{P(y_{i1}=0)P(y_{i2}=1)}{P(T(\vec{y_i})=1)}\\
&= \frac{\left(1-\pi\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)\right)\pi\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}{P(y_{i1}+y_{i2}=1\ )}\\
&= \frac{\left(1-\pi\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)\right)\pi\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}{P(y_{i1}=0\ )P(y_{i2}=1\ )P(y_{i1}=1\ )P(y_{i2}=0\ )}\\
&\frac{\left(1-\pi\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)\right)\pi\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}{\left(1-\pi\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)\right)\pi\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)+\pi\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)\left(1-\pi\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)\right)}
\end{aligned}
\tag{3.11}
$$

$$
\begin{aligned}
&= \frac{\left(1-\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)}{\left(1-\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)+\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(1-\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)} \\
&= \frac{\left(\frac{1+e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})-1}}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)}{\left(\frac{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)-1}}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)+\left(\frac{1}{1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}\right)\left(\frac{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)-1}}{1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right)} \\
&= \frac{\left(\frac{e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}}{\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)}\right)}{\left(\frac{e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}}{\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)}\right)+\left(\frac{e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}{\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)}\right)} \\
&= \frac{\left(\frac{e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}}{\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)}\right)}{\left(\frac{e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}{\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)}\right)} \\
&= \left(\frac{e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}\left(1+e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}\right)\left(1+e^{-(\alpha_i+\vec{x}'_{i2}\vec{\beta})}\right)}{\left(1+e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}\right)\left(1+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)\left(e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}\right)}\right) \\
&= \left(\frac{e^{-(\alpha_i+\vec{x}'_{i1}\vec{\beta})}}{e^{-\left(\alpha_i+\vec{x}'_{i1}\vec{\beta}\right)}+e^{-\left(\alpha_i+\vec{x}'_{i2}\vec{\beta}\right)}}\right) \\
&= \frac{e^{-\alpha_i}e^{-(\vec{x}'_{i1}\vec{\beta})}}{e^{-\alpha_i}\left(e^{-\left(\vec{x}'_{i1}\vec{\beta}\right)}+e^{-\left(\vec{x}'_{i2}\vec{\beta}\right)}\right)} \\
&= \frac{e^{-(\vec{x}'_{i1}\vec{\beta})}}{\left(e^{-\left(\vec{x}'_{i1}\vec{\beta}\right)}+e^{-\left(\vec{x}'_{i2}\vec{\beta}\right)}\right)} \\
&= \frac{e^{-(\vec{x}'_{i1}\vec{\beta})}\left(e^{\vec{x}'_{i2}\vec{\beta}}e^{\vec{x}'_{i1}\vec{\beta}}\right)}{\left(e^{-\left(\vec{x}'_{i1}\vec{\beta}\right)}+e^{-\left(\vec{x}'_{i2}\vec{\beta}\right)}\right)\left(e^{\vec{x}'_{i2}\vec{\beta}}e^{\vec{x}'_{i1}\vec{\beta}}\right)} \\
&= \frac{e^{\vec{x}'_{i2}\vec{\beta}}}{e^{\vec{x}'_{i1}\vec{\beta}}+e^{\vec{x}'_{i2}\vec{\beta}}}
\end{aligned}
\tag{3.12}
$$

La cual coincide con la función de verosimilitud del modelo de regresión de Cox estratificado. Puesto que como ya habíamos mencionado anteriormente, los coeficientes β de cada covariable no cambian a través de los estratos, por lo tanto estaríamos comprobando la afirmación hecha por Gauderman (1999), en la cual una razón hazard calculada utilizando un modelo de regresión de Cox estratificado, se puede interpretar de la misma forma que una razón de odds, obtenida al aplicar la regresión logística condicional. Sin embargo, no necesariamente se cumple que las funciones de verosimilitud sean iguales, para el modelo de regresión logística condicional y el modelo de regresión de Cox estratificado, cuando el pareamiento se hace $1 : n$, con $n > 1$.

3.1. Ventajas y Desventajas

Una ventaja que presenta el uso de la regresión logística condicional, consiste en que no es necesario asumir riesgos proporcionales dentro del estrato, y la razón de odds se puede interpretar de la manera usual, sin embargo, si queremos interpretar la razón de odds como una razón hazard, es necesario asumir la existencia de riesgos proporcionales dentro del estrato.

Por otra parte, resulta ser una desventaja el emplear la regresión logística condicional, cuando se conoce la edad de aparición de la enfermedad en los casos y se tienen hermanos que podrían ser catalogados como controles, pero que debido a que estos posibles controles, en la actualidad presentan la enfermedad bajo estudio, son descartados en la regresión logística condicional bajo el diseño de Gauderman, sin tener en consideración que se encontraran aliviados a la edad en la que el caso se enfermo, contrario a lo que sucede cuando se emplea el modelo de regresión de Cox estratificado, quien admite este tipo de controles dentro del análisis, lo que conlleva a contar con mayor número de personas.

3.2. Similitudes

Tanto la regresión logística condicional, como el modelo de regresión de Cox estratificado, utilizan métodos de maxima verosimilitud para calcular los parámetros, Ambas metodologías sirven para controlar el efecto de la confusión de alguna variable.

CAPÍTULO 4

Métodos

4.1. Introducción

Las enfermedades autoinmunes están conformadas por un conjunto de condiciones crónicas caracterizadas por una pérdida de la tolerancia inmunológica hacia antígenos propios, y conforman un grupo heterogéneo de desordenes en los que, dados múltiples alteraciones en el sistema inmune, se desencadena un espectro de síndromes que afectan ciertos órganos en forma específica o en forma sistémica (Anaya, Shoenfeld, Correa, Garcia, Carraso y Cervera, 2005). Una condición epidemiológica que orienta en el interés de realizar el estudio en familias es que una característica que presentan las enfermedades autoinmunes complejas consiste en que los individuos que se encuentran afectados tienden a agruparse en familias (agregación familiar, también conocido como riesgo de ocurrencia ó λ). La agregación familiar de un

fenotipo ocurre cuando el fenotipo se presenta con una frecuencia mayor en familias de un individuo afectado, que la frecuencia esperada en la población general.
Es objeto de interés es estudiar el genoma de los individuos que padecen de MAS y sus familiares (hermanos (as), primos (as)). El Síndrome de Autoinmunidad Múltiple (MAS), fue descrito inicialmente por Humbert y Dupond en 1988 como un síndrome que consiste en la presencia de tres o mas enfermedades autoinmunes en un solo paciente (Humbert y Dupond, 1988). Se estima que la incidencia de las enfermedades autoinmunes es encuentra alrededor de 90 por 100000 habitantes por año, y su prevalencia es del 3 % , de la población Norteamericana (Cooper y Stroehla, 2003). Por lo tanto la presencia de SEMIMAS en los individuos en estudio se define como la presencia de dos enfermedades autoinmunes, este fenotipo también representa interés dentro del presente estudio.

4.2. Generalidades

Para el análisis se contó con la información suministrada por el Dr. Juan Manuel Anaya, la cual consistía en una base de datos en donde se recopilaron datos referentes a pacientes y sus familiares que presentaban MAS o SEMIMAS. Algunas de las variables suministradas más relevantes fueron las siguientes: 786 alelos correspondientes a 393 microsatélites. (Ver Apéndice B. Lista de Marcadores Genéticos) Un código, que identifica la familia. El sexo que tiene cada persona del estudio, codificando con 0 a los hombres y 1 a las mujeres. Esta variable es de interés porque los estudios epidemiológicos muestran que las mujeres tienden a ser más susceptibles que los hombres para desarrollar enfermedades autoinmunes. Así, esta es una potencial variable de confusión. Se considera que una persona sufre de MAS si tiene al menos tres de las siguientes enfermedades (o que sufre de SEMIMAS si presenta dos de las siguientes enfermedades):

Diabetes Mellitus tipo 1 (T1D), Lupus Eritematoso Sistémico (SLE), Síndrome Antifosfolípido (APS), Artritis Reumatoide (RA), Síndrome de Sjögren (SS), Miastemia Gravis (MG), Vasculitis (Churg-Straus, Vasculitis Cutanea, Poliangeitis microscópica, Crioglobulinemia, Arteritis de Células Gigantes) (V), Escleroderma (SSc), Dermato-polimiositis (DPM), Enfermedad inflamatoria intestina: Colitis ulcerativa y Enfermedad de Crohn's (IBD), Anemia Perniciosa (PA), Enfermedad Tiroidea autoinmune (AITD), Enfermedad Celiaca (CD), Artritis Juvenil (JRA), Vitíligo (VIT), Enfermedades inflamatorias biliares: Cirrosis Biliar Primaria y Conlangitis esclerosante (IBDS), Hepatitis Autoinmune (AH), Enfermedades Autoinmunes Desmielinizantes: Mielitis transversa y Esclerosis Múltiple (DAD), Policondritis Recurrente (RP), Enfermedad de Addison (AD), Glomerulonefritis (GN), Citopenias Autoinmunes: AHA y PTI (AC).

4.3. Metodología

En esta sección del trabajo se presenta la aplicación de los métodos consignados en el marco teórico, utilizando la información suministrada por el grupo de investigación de la unidad de Biología Celular e Inmunogenética de la Corporación para Investigaciones Biológicas "CIB". Con el objetivo de calcular el tamaño del efecto de los marcadores genéticos sobre el riesgo de tener MAS o SEMIMAS, se usaron las herramientas del modelo de regresión de Cox estratificado y la regresión logística condicional.

4.3.1. Muestra de estudio

La muestra de estudio, estuvo conformada por personas de nacionalidad colombiana que acudían a consulta médica con el Dr. reumatólogo Juan Manuel Anaya. Quienes

asistían por sospechas de la presencia de por lo menos una enfermedad autoinmune, debido al cuadro clínico que estas expresaban. Una vez que el Dr. Anaya, confirma las sospechas de la presencia de MAS en los pacientes, es decir, que verifica médicamente por medio de exámenes de laboratorio la presencia de ciertos antígenos, que permiten la comprobación de cada una de las enfermedades autoinmunes mencionadas anteriormente, él y su equipo de trabajo bajo el consentimiento del propio paciente, hace una recopilación de la historia familiar de éste, en donde se revisan los familiares en primer y segundo grado de consanguinidad del mismo. Permitiendo así, conocer si los familiares del paciente con MAS (denominado caso), también padecen alguna enfermedad autoinmune. Dado que se tratan de enfermedades de baja prevalencia en la población, la recolección de la información fué una tarea ardua, que contó de gran persistencia y dedicación por parte del Dr. Anaya y su equipo de colaboradores, debido a que les tomó cerca de 7 años el conseguir la información, con la que hoy se cuenta, es por esto, que esta muestra es un recurso muy valioso, para entender con ayuda de herramientas estadísticas el efecto que existe entre algunos alelos y el desarrollo de la enfermedad. Posteriormente, a cada uno de los pacientes que cumplían con el criterio de tener MAS, junto con los familiares, se les tomaron muestras de sangre, a las cuales se les realizó extracción de ADN, en el laboratorio de la Unidad de Biología Celular e Inmunogenética de la CIB, el cual fué enviado al "Center For Medical Genetics, Marshfield Wisconsin", con el fin de realizarles un "Genome Scan", el cual permitió tener la información genética para cada uno de los marcadores de los cromosomas. Lo que nos permitió contar para la aplicación del modelo de regresión de Cox estratificado, con un total de 19 familias a ser estudiadas, las cuales están conformadas por 68 personas, de las cuales 53 eran mujeres, ver Tabla C,1. Para la aplicación de la regresión logística condicional, se tuvo un total de 19 familias a ser estudiadas, las cuales están conformadas por 58 personas, de las cuales 45 eran mujeres, ver Tabla C,2. Por otra parte, para aquellas personas que

presentaran SEMIMAS, es decir aquellas personas que tuvieran dos enfermedades autoinmunes, para la aplicación del modelo de regresión de Cox estratificado, se investigaron un total de 31 familias, las cuales están conformadas por 119 personas, de las cuales 91 eran mujeres, ver Tabla C,3. Finalmente, para la aplicación de la regresión logística condicional, en quienes padecen de SEMIMAS, se trabajó con un total de 31 familias, las cuales están conformadas por 99 personas, de las cuales 79 eran mujeres, ver Tabla C,4.

4.3.2. Recolección y almacenamiento de la muestra

La toma de muestra sanguínea se realizó por punción de la vena antecubital, con aguja vacuitainer, con el fin de recolectar 20 cc de sangre periférica, distribuida en dos tubos Becton Dickenson®:

- Tubo tapa lila con EDTA, donde se almacenaron $4cc$ de sangre total, para la extracción de DNA ($200\mu l$). La restante será almacenado en congelador a -70^oC.
- Tubo tapa roja sin anticoagulante, en donde se colocaron $7cc$ de sangre total, para obtener suero, el cual permitirá la determinación de anticuerpos confirmatorios de enfermedad autoinmne. El restante será almacenado en congelador a -70^oC.

Posteriormente las muestras son separadas en suero y sangre total en volúmenes de $0{,}2ml$. La separación de los componentes se hizo en cabina de flujo laminar y las muestras son rotuladas con el código asignado a cada participante para ser congeladas a -70^oC.

Previo a su almacenamiento quedará registro de la fecha, número de alícuotas y

persona responsable del procedimiento, esto para iniciar la cadena de custodia de las muestras bajo la responsabilidad del grupo de investigación.

4.3.3. Análisis Bioquímico y Genético

El análisis bioquímoco, se realiza para confirmar algún diagnóstico de tipo autoinmune; este proceso se lleva acabo mediante la prueba de inmuno-absorción a enzimas (enzymelinked immunosorbent assay) ELISA, la cual es una técnica sensible, versátil, precisa, reproducible, de carácter cuantitativo y cualitativo, que ayuda a la determinación de antígenos (Ag) o anticuerpos (Ac) en una muestra biológica.
La ELISA es un inmunoensayo ampliamente utilizado como herramienta diagnostica y de investigación biológica. El principio básico de la técnica de ELISA es el duo de una Ag o Ac conjugado con una enzima, la cual es capaz de reaccionar con su sustrato, generando una reacción de color donde se produce la interacción inmunológica antígeno-anticuerpo. El cambio de color es monitoreado visualmente (cualitativo) o por el uso de espectrofotometría (cuantitativo), para determinar la cantidad de analito presente en la muestra.
Un paso esencial, en este tipo de pruebas es la separación de la enzima marcada unida durante la reacción y el marcaje libre o inespecífico que se genera durante la prueba. De igual manera, en el caso de determinación de anticuerpos, puede darse, incluso discriminación y cuantificación de isótopos, dependiendo de la especificidad del antígeno utilizado.

Extracción de ADN

La extracción de ADN se realiza con QIAamp DNA Blood Minikit (QUIAGEN, Germany) a partir de una muestra de $200\mu l$ de sangre, según las instrucciones

del fabricante. El DNA obtenido será verificado por medio de espectrofotometría (GENESIS II, USA).

Genotipificación

El principio de ensayo para PCR por INNO-LiPA HLA-A Multiplex, Innogenetics 25011 v1: La muestra de ADN a amplificar mediante PCR se introduce en una mezcla de reactivos que contiene un tampón con un exceso de desoxinucleosido 5-trifosfatos (dNTPs), "primers" (oligonucleotidos cebadores) biotinilados, y ADN polimerasa termoestable. Los primers amplifican la secuencia diana del ADN. Las dos cadenas de la hélice de ADN se separan (desnaturalizando) por calentamiento, exponiendo las secuencias diana a los primers. Tras enfriar la mezcla a una temperatura concreta, estos primers se ligan a regiones complementarias de secuencias que flanquean a la secuencia diana (anillamiento). A otra temperatura concreta, la ADN polimerasa termoestable utiliza el exceso de dNTPs, extendiendo los primers anillados a lo largo del ADN molde diana (extensión). De esta forma, tras un ciclo se obtiene dos copias exactas, biotiniladas, de la secuencia diana. Tras varios ciclos se obtienen dos copias exactas, biotilinadas, de la secuencia diana. Tras varios ciclos se obtiene un número mayor de copias biotiniladas de la secuencia diana. Los principios de ensayo para el HLA por INNO-LiPA HLA-A Update, Innogenetics 25003 v3: Las pruebas de tiraje INNO-LIPA HLA se basan en los principios de hibridación reversa que se resumen en la figura 1. el material de ADN biotilinado amplificado se desnaturaliza químicamente, y las hebras separadas se hibridan con sondas de oligonecleotidos específicos, inmovilizadas en líneas paralelas sobre tiras basadas en membranas. Esto va seguido de una fase de lavado astrigente a fin de eliminar cualquier material amplificado incorrectamente emparejado. Tras el lavado astrigente, se añade estreptavidina conjugada con fosfatasa alcalina, que queda ligada a cualquier

hibrido biotinilado que se haya formado con anterioridad. La incubación con una solución sustrato que contiene un cromógeno produce un precipitado de color púrpura/marrón. La reacción se interrumpe mediante una fase de lavado, tras la que se registra el patrón de reactividad de las sondas. Posteriormente, los productos de la amplificación se hibridan utilizando 2 tiras de tiraje que llevan fijadas 43 sondas especificas de secuencia, así como 2 líneas de control.

4.3.4. Criterios de inclusión para MAS y SEMIMAS

Para que una persona hiciera parte de la muestra, es necesario que presente tres enfermedades autoinmunes, para que sea clasificada como un caso, dentro del estudio de MAS, el cual es el criterio médico para tener esta enfermedad, mientras que para los casos correspondientes a la muestra de SEMIMAS, se establece que los pacientes deben sufrir de dos enfermedades autoinmunes; adicionalemente se consideró necesario que cada paciente tuviera como mínimo un hermano(a) o primo(a), que se encontrara vivo, con el fin de tomarlo como un control, lo cual permite controlar algunos de los factores ambientales debido a que se encuentran expuestos a ambientes similares, al pertenecer a la misma familia además se controlan condiciones genéticas al igual que la raza, por compartir un grado de consanguinidad tan cercano. Finalmente, la persona elegida como control, se debe encontrar libre de la enfermedad (MAS o SEMIMAS) a la edad en la que el caso adquirió la enfermedad (esto para la aplicación de la regresión logística condicional).

4.3.5. Comparación información de modelos para MAS y SEMIMAS

Como se observa en las Tablas de la C.1 a la C.4 (Ver Apéndice C. Tablas Descriptivas Familias Seleccionadas), el número de personas seleccionadas según los criterios establecidos anteriormente, varia dependiendo de la metodología que se emplee, es decir, entre el modelo de regresión logística condicional o el modelo de regresión de Cox estratificado, tanto para el estudio de MAS como SEMIMAS, este cambio se refleja en el número de personas por familia que entran en el estudio. Además, se aprecia que siempre se trabaja con el mismo número de familias lo que nos permite comparar los resultados obtenidos a pesar de la variación que existe en el número de personas seleccionadas; para el estudio de MAS se cuenta con un total de 19 familias, en donde se tiene un total de 68 personas seleccionadas para la aplicación de la regresión de Cox estratificado y 58 para la aplicación de la regresión logística condicional, para el estudio de SEMIMAS se trabajó con 31 familias, dentro de las cuales se tienen 119 personas, para la aplicación de la regresión de Cox estratificado y 96 para la regresión logística condicional. Finalmente, se aprecia la existencia de más mujeres que hombres dentro de la muestra tanto para MAS como SEMIMAS, debido principalmente a que las enfermedades autoinmunes se consideran más comunes en las mujeres que en los hombres, por la predisposición genética propia del genéro.

Tabla 4.1: Comparación información de modelos para MAS

Metodología Análisis	**Edad índice promedio**	**N° Familias**	**Total caso-control**	**N° Mujeres**
Modelo de regresión de Cox estratificado	32.25 ±13,206	19	68	53
Regresión logística condicional	31.63 ±13,267	19	58	45

Tabla 4.2: Comparación información de modelos para SEMIMAS

Metodología Análisis	**Edad índice promedio**	**N° Familias**	**Total caso-control**	**N° Mujeres**
Modelo de regresión de Cox estratificado	31.97 ±12,87	31	119	91
Regresión logística condicional	31.52 ±13,092	31	96	76

Las Tablas 4,1 y 4,2, permiten ver como la edad índice, la cual se encuentra definida como la edad en la que el caso adquirió la enfermedad, son muy similares, pero sin embargo no son iguales como se creía inicialmente, debido a que cuando se trabaja con el modelo de regresión de Cox estratificado, en la selección de la muestra se admiten controles que puede que en estos momentos se encuentren enfermos, pero que sin embargo a la edad en la que el caso se enfermo estaban aliviados, por lo tanto la edad índice promedio para las personas que se utilizan para la aplicación de la metodología de Cox estratificado son un poco mayores.

4.3.6. Información Genética

La información genética recopilada se encuentra desagregada por Alelos los cuales a su vez conforman los marcadores, los datos que se poseen por cada uno de los alelos se encuentran expresados como un peso molecular en pares de bases (En genética un par de bases consiste en dos nucleótidos opuestos y complementarios en las cadenas

de ADN y ARN que están conectadas por puentes de hidrógeno. En el ADN adenina y timina así como guanina y citosina, pueden formar un par de bases).

4.3.7. Categorización

Se estableció que un alelo es catalogado como largo si presenta un peso molecular mayor o igual que el valor de la mediana de todos los pesos moleculares observados para dicho marcador, dentro de la población en estudio. En caso contrario, el alelo se consideró como corto.

Para realizar el análisis de cada uno de los marcadores, fue necesario definir la siguiente categorización: Si dentro de un marcador, los alelos que lo componen son clasificados como cortos, entonces se le asignará un valor de cero, en caso contrario es decir, que los alelos sean denominados alelos largos, se le asignará un valor de dos. Finalmente, en el caso en el que los alelos del marcador sea uno corto y uno largo, o al contrario, el valor asignado para cada una de estas posibilidades sera uno.

Para aquellos casos especiales en que la mediana de los marcadores no coincidieran dentro de la información disponible para aplicar la regresión logística condicional y la regresión de Cox, se utilizó como medida de comparación la moda. Esto con el fin de hacer comparables los resultados que se obtienen al aplicar cada una de las metodologías mencionadas anteriormente.

Dicotomización

Una vez establecida la categorización de cada uno de los marcadores, se tienen 3 posibles resultados los cuales son 0, 1 y 2, por lo tanto se definió una dicotomización $(0, 1)$, que permitiera aplicar las metodologías establecidas (regresión logística condicional y el modelo de regresión de Cox estratificado), generando 6 diferentes posibilidades, de la siguiente manera a dos de las categorías se les asigna el número uno y a la restante el número cero, este procedimiento se hizo en cada una de las 6

posibilidades.

Análisis para SEMIMAS

Tabla 4.3: Marcadores que estuvieron significativamente asociados con SEMIMAS, de acuerdo a la regresión de Cox estratificado o la regresión logística condicional. Para propósitos de comparación también se incluyen los resultados de estos marcadores, obtenidos en el análisis de MAS

	SEMIMAS				**MAS**			
	Cox		**Condicional**		**Cox**		**Condicional**	
Marcador	**Valor p**	**Razón Hazard**	**Valor p**	**Razón de riesgo**	**Valor p**	**Razón Hazard**	**Valor p**	**Razón de riesgo**
$ATTT030^{e}$	**0.0500**	**0.094**	**0.0514**	**0.097**	0.2246	0.191	0.2334	0.198
$ATTT030^{f}$	**0.0500**	**10.615**	**0.0514**	**10.354**	0.2246	5.243	0.2334	5.055
$GATA11C06N^{g}$	**0.0490**	**0.120**	0.0740	0.141	0.9967	0.000	0.9971	0.000
$GATA12H10^{i}$	**0.0399**	**0.107**	**0.0546**	**0.121**	**0.0546**	**0.121**	0.0767	0.140
$GATA12H10^{j}$	**0.0399**	**9.319**	**0.0546**	**8.248**	**0.0546**	**8.248**	0.0767	7.168
$GATA21F05^{k}$	**0.0578**	**4.828**	**0.0578**	**4.828**	0.1633	3.310	0.1633	3.310
$GATA21F05^{l}$	**0.0578**	**0.207**	**0.0578**	**0.207**	0.1633	0.302	0.1633	0.302

$GATA65C03M^{m}$	**0.0570**	**3.570**	**0.0280**	**6.060**	**0.0357**	**9.736**	**0.0333**	**11.948**
$GATA65C03M^{n}$	**0.0570**	**7.841**	**0.0280**	**0.165**	**0.0357**	**0.103**	**0.0333**	**0.084**
$GATA65C03M^{\tilde{n}}$	0.1208	0.354	**0.0577**	**0.212**	0.0728	0.143	0.0641	0.116
$GATA65C03M^{o}$	0.1208	2.824	**0.0577**	**4.728**	0.0728	6.990	0.0641	8.658
$GATA68F07^{p}$	**0.0599**	**0.127**	**0.0564**	**0.119**	0.0725	0.137	0.0681	0.128
$GATA68F07^{q}$	**0.0599**	**7.880**	**0.0564**	**8.435**	0.0725	7.310	0.0681	7.841
$GATA68F07^{r}$	**0.0599**	**7.880**	**0.0564**	**8.435**	0.0725	7.310	0.0681	7.841
$GATA68F07^{s}$	**0.0599**	**0.127**	**0.0564**	**0.119**	0.0725	0.137	0.0681	0.128
$GATA70E11^{t}$	**0.0298**	**0.075**	**0.0395**	**0.084**	0.0964	0.124	0.1430	0.151
$GATA70E11^{u}$	**0.0298**	**13.251**	**0.0395**	**11.925**	0.0964	8.055	0.1430	6.631
$GATA70E11^{v}$	**0.0403**	**12.443**	**0.0403**	**12.443**	0.1006	8.354	0.1006	8.354
$GATA70E11^{w}$	**0.0403**	**0.080**	**0.0403**	**0.080**	0.1006	0.120	0.1006	0.120
$GGAA20G04^{x}$	**0.0558**	**0.125**	0.1150	0.171	**0.0409**	**0.104**	**0.0575**	**0.117**
$GGAA20G04^{y}$	**0.0558**	**7.983**	0.1150	5.862	**0.0409**	**9.646**	**0.0575**	**8.521**
$GGAA6D03N^{z}$	**0.0302**	**11.778**	**0.0316**	**11.442**	0.1054	7.240	0.1113	6.961
$GGAA6D03N^{aa}$	**0.0302**	**0.085**	**0.0316**	**0.087**	0.1054	0.138	0.1113	0.144
$SraP^{ab}$	0.0971	0.332	**0.0568**	**0.225**	0.1830	0.402	0.1045	0.273
$SraP^{ac}$	0.0971	3.012	**0.0568**	**4.442**	0.1830	2.487	0.1045	3.663

Tabla 4.4: Tabla comparativa entre el Modelo de Regresión de Cox estratificado y la Regresión Logística Condicional

	SEMIMAS		**MAS**	
Marcador	**Cox**	**Condicional**	**Cox**	**Condicional**
	Valor p	**Valor p**	**Valor p**	**Valor p**
$ATTT030^{e}$	+	+	-	-
$ATTT030^{f}$	+	+	-	-
$GATA11C06N^{g}$	+	-	-	-
$GATA12H10^{i}$	+	+	+	-
$GATA12H10^{j}$	+	+	+	-
$GATA21F05^{k}$	+	+	-	-
$GATA21F05^{l}$	+	+	-	-
$GATA65C03M^{m}$	+	+	+	+
$GATA65C03M^{n}$	+	+	+	+
$GATA65C03M^{\tilde{n}}$	-	+	-	-
$GATA65C03M^{o}$	-	+	-	-
$GATA68F07^{p}$	+	+	-	-
$GATA68F07^{q}$	+	+	-	-
$GATA68F07^{r}$	+	+	-	-
$GATA68F07^{s}$	+	+	-	-
$GATA70E11^{t}$	+	+	-	-
$GATA70E11^{u}$	+	+	-	-
$GATA70E11^{v}$	+	+	-	-
$GATA70E11^{w}$	+	+	-	-
$GGAA20G04^{x}$	+	-	+	+
$GGAA20G04^{y}$	+	-	+	+
$GGAA6D03N^{z}$	+	+	-	-
$GGAA6D03N^{aa}$	+	+	-	-
$SraP^{ab}$	-	+	-	-
$SraP^{ac}$	-	+	-	-

* Las casillas que se encuentran con signo más representa los que son significativos después de aplicar el Modelo de Regresión de Cox estratificado y la Regresión Logística Condicional.

** Las casillas que se encuentran con signo menos representan los que no son significativos después de aplicar el Modelo de Regresión de Cox estratificado y la

Regresión Logística Condicional.

[e] La variable dicotómica fue definida como cero si el individuo era homocigoto para alelo corto, uno en otro caso. (Alelo corto:≥128 pares de bases).

[f] La variable dicotómica fue definida como uno si el individuo era homocigoto para alelo corto, cero en otro caso. (Alelo corto: ≥128 pares de bases).

[g] La variable dicotómica fue definida como cero si el individuo era homocigoto para alelo largo, uno en otro caso. (Alelo largo: ≥160 pares de bases).

[i] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥251 pares de bases).

[j] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo:≥251 pares de bases).

[k] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥259 pares de bases).

[l] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: ≥259 pares de bases).

[m] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥126 pares de bases).

[n] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: ≥126 pares de bases).

[ñ] La variable dicotómica fue definida como cero si el individuo era homocigoto para alelo largo, uno en otro caso. (Alelo largo: ≥126 pares de bases).

[o] La variable dicotómica fue definida como uno si el individuo era homocigoto para alelo largo, cero en otro caso. (Alelo largo: ≥126 pares de bases).

[p] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥159 pares de bases).

[q] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: ≥159 pares de bases).

[r] La variable dicotómica fue definida como cero si el individuo era homocigoto para alelo largo, uno en otro caso. (Alelo largo: ≥159 pares de bases).

[s] La variable dicotómica fue definida como cero si el individuo era homocigoto para alelo largo, uno en otro caso. (Alelo largo: ≥159 pares de bases).

[t] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥230 pares de bases).

[u] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: ≥230 pares de bases).

[v] La variable dicotómica fue definida como cero si el individuo era homocigoto para alelo largo, uno en otro caso. (Alelo largo: ≥230 pares de bases).

[w] La variable dicotómica fue definida como uno si el individuo era homocigoto para alelo largo, cero en otro caso. (Alelo largo: ≥230 pares de bases).

[x] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥149 pares de bases).

[y] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: ≥149 pares de bases).

[z] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥138 pares de bases).

[aa] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: ≥138 pares de bases).

[ab] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥128 pares de bases).

[ac] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: $\geq$128 pares de bases).

Todas las razones de riesgo y los valores p están ajustados por genero.

Tabla 4.5: Marcadores que estuvieron significativamente asociados con MAS, de acuerdo a la regresión de Cox estratificado o la regresión logística condicional. Para propósitos de comparación también se incluyen los resultados de estos marcadores, obtenidos en el análisis de SEMIMAS

	MAS				**SEMIMAS**			
	Cox		**Condicional**		**Cox**		**Condicional**	
Marcador	**Valor p**	**Razón Hazard**	**Valor p**	**Razón de riesgo**	**Valor p**	**Razón Hazard**	**Valor p**	**Razón de riesgo**
GATA12H10[a]	**0.0546**	**0.121**	0.0767	0.140	**0.0399**	**0.107**	**0.0546**	**0.121**
GATA12H10[b]	**0.0546**	**8.248**	0.0767	7.168	**0.0399**	**9.319**	**0.0546**	**8.248**
GATA65C03M[c]	**0.0357**	**9.736**	**0.0333**	**11.948**	**0.0570**	**3.570**	**0.0280**	**6.060**
GATA65C03M[d]	**0.0357**	**0.103**	**0.0333**	**0.084**	**0.0570**	**7.841**	**0.0280**	**0.165**
GGAA20G04[e]	**0.0409**	**0.104**	**0.0575**	**0.117**	**0.0558**	**0.125**	0.1150	0.171
GGAA20G04[f]	**0.0409**	**9.646**	**0.0575**	**8.521**	**0.0558**	**7.983**	0.1150	5.862

Tabla 4.6: Tabla comparativa entre el Modelo de Regresión de Cox estratificado y La Regresión Logística Condicional

	MAS		**SEMIMAS**	
Marcador	**Cox**	**Condicional**	**Cox**	**Condicional**
	Valor p	**Valor p**	**Valor p**	**Valor p**
GATA12H10[a]	+	-	+	+
GATA12H10[b]	+	-	+	+
GATA65C03M[c]	+	+	+	+
GATA65C03M[d]	+	+	+	+
GGAA20G04[e]	+	+	+	-
GGAA20G04[f]	+	+	+	-

* Las casillas que se encuentran con signo más representa los que son significativos después de aplicar el Modelo De Regresión De Cox y La Regresión Logística Condicional.

** Las casillas que se encuentran con signo menos representan los que no son significativos después de aplicar el Modelo De Regresión De Cox y La Regresión Logística Condicional.

[a] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: $\geq$259 pares de bases).

[b] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: $\geq$126 pares de bases).

[c] La variable dicotómica fue definida como uno si el individuo era homocigoto para alelo largo, cero en otro caso. (Alelo largo: $\geq$126 pares de bases).

[d] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno

en otro caso. (Alelo: ≥159 pares de bases).

[e] La variable dicotómica fue definida como uno si el individuo era heterocigoto, cero en otro caso. (Alelo: ≥138 pares de bases).

[f] La variable dicotómica fue definida como cero si el individuo era heterocigoto, uno en otro caso. (Alelo: ≥128 pares de bases).

Todas las razones de riesgo y los valores p están ajustados por genero.

4.3.8. Algunas Interpretaciones de los resultados

Para el estudio de MAS, en la Tabla 4,6, se observa que después de aplicar la regresión logística condicional y el modelo de regresión de Cox estratificado, existen 3 marcadores genéticos que resultaron estar asociados de manera significativa con la enfermedad, los cuales son (GATA12H10, GATA65C03M, GGAA20G04), en donde cada uno de ellos fue dicotómizado de dos maneras distintas, por ejemplo para el marcador GATA12H10 se tiene que fue significativo cuando la variable dicotómica, se definió como uno, si el individuo era heterocigoto para el Alelo: ≥259 pares de bases, al igual que para el Alelo: ≥126 pares de bases; Al comparar las metodologías se puede observar que existen más resultados significativos cuando se emplea el modelo de regresión de Cox estratificado, que la regresión logística condicional para el caso específico del estudio de MAS, sin embargo al revisar la Tabla 4,5, se aprecia que la razón de riesgo para cada una de las metodologías, presenta el mismo sentido, es decir, que ambas metodologías apuntan a resultados muy similares en la cuantificación del tamaño de efecto, que puede tener un marcador especifico en el padecimiento de la enfermedad (MAS); cabe anotar que cada uno de los marcadores genéticos, que fueron anteriormente presentados han sido ajustados por el género, el cual para su dicotomización se estableció como uno para las mujeres y cero para los hombres, teniendo en cuenta las consideraciones genéticas de predisposición a la

enfermedad, la forma de pareamiento que se utiliza en este trabajo, es uno variable "1:n", lo que significa que por cada uno de los casos seleccionados se puede tener más de un control, aquí los estratos corresponden a cada una de las familias que conforman el análisis. Para ilustrar las metodologías aplicadas, a continuación se presenta un modelo encontrado por cada metodología: En el modelo de regresión de Cox estratificado, se tiene:

$$\begin{aligned} h_k(t) &= h_{0k}(t)\exp(\beta_1 X_1 + \beta_2 X_2) \quad \text{donde} \quad k = 1, 2, \ldots, 19 \\ &= h_{0g}(t)\exp(2{,}27579\,(\text{GATA65C03M}^{\text{c}}) + 3{,}03620\,(\text{Sexo})\,) \end{aligned}$$

El hazard de adquirir MAS, en un momento dado del periodo de seguimiento, para un paciente de sexo masculino que presenta el marcador genético GATA65C03M es 9,736 (Ver Tabla 4,5) veces más alto que para un sujeto que no tiene el marcador genético GATA65C03M, que haya sido observado la misma cantidad de tiempo.

Ahora analizando los resultados obtenidos al aplicar la metodología de la regresión logística condicional, se tiene:

$$\begin{aligned} \text{logit}(\psi_i) &= \exp(\beta_1 X_1 + \beta_2 X_2) \\ &= \exp(2{,}48054\,(\text{GATA65C03M}^{\text{c}}) + 2{,}48671\,(\text{Sexo})\,) \end{aligned}$$

El riesgo relativo de adquirir MAS, para un paciente de sexo masculino que presenta el marcador genético GATA65C03M es 11,948 (Ver Tabla 4,5) veces más alto que para un sujeto que no tiene el marcador genético GATA65C03M.

CAPÍTULO 5

Conclusiones

- La afirmación hecha por Gauderman y colaboradores, a saber, que la razón de odds correspondiente a una variable genética dicotómica calculada a partir de un modelo de regresión logística condicional puede interpretarse como una razón hazard, cuando los datos provienen de un estudio caso control en donde se controla el efecto confusor de la edad de aparición de la enfermedad, resulta ser cierta. Sin embargo, es necesario asumir riesgos proporcionales dentro de cada familia. Asi, Gauderman y colaboradores justificaron incorrectamente su afirmación, utilizando la idea errónea de que la función de verosimilitud del modelo de regresión de Cox coincide con la función de verosimilitud del modelo de regresión logística condicional. En realidad, no es la función de verosimilitud del modelo de regresión de Cox, sino la función de verosimilitud del modelo de regresión de Cox estratificado, la que coincide con la función de verosimilitud

de la regresión logística condicional.

- Cuando un caso tiene más de un pariente control y uno de los controles ha manifestado la enfermedad, la regresión logística condicional ignora la edad de aparición de la enfermedad del control; esto puede implicar una pérdida importante de información. Además, si el modelo de Cox es adecuado, el descartar hermanos que se hayan enfermado antes de la edad índice es un desperdicio de información, ya que, en caso de que se haya examinado este tipo de hermanos, sería más recomendable usar regresión de Cox estratificada para analizar los datos, incorporando en el análisis la edad de aparición de la enfermedad tanto de los casos como de los controles. Por lo tanto, en las dos situaciones anteriores, el uso de un modelo estratificado de Cox en el que se modele explícitamente el efecto del factor genético sobre la edad de aparición de la enfermedad puede ser más razonable. Así, cuando el modelo de Cox estratificado es adecuado, la regresión de Cox podría ser una mejor estrategia de análisis que la misma regresión logística condicional.

- Cuando no se cumple el supuesto de riesgos proporcionales dentro de cada estrato (familia), la regresión logística condicional, es más adecuada que el modelo de regresión de Cox estratificado. Sin embargo, en esta situación, no es posible usar controles que se hayan enfermado antes de la edad índice, y la interpretación de la razón de odds como una razón de hazards ya no es válida.

Bibliografía

[1] Anaya JM, Shoenfeld Y, Correa PA, Garcia-Carraso M,Cervera R. (2005). "Autoimmunity and Autoimmune Disease.".Medellín, CIB.

[2] Cooper GS, Stroehla BC. "The epidemiology of autimmune deseases" *Autoimmun Rev 2003*. 119-125.

[3] Duncan,T. (2004). "*Statistical Methods in Genetic Epidemiology*". New York, Oxford University Press, Inc.

[4] Elston,R. Olson,J. Palmer,L. (2002). "*Bioestadistical Genetics and Genetic Epidemiology*".

[5] Frees,E. (2004).“*Longuitudinal and Panel Data*”. Cambridge.

[6] Gauderman, J. Witte,J. Thomas,D “Family-Based Association Studies” *Journal Of The National Cancer Institute Monographs*, (1999).

[7] Guerrero,R. González,C. Medina,E. (1981). “*Epidemiología*”.

[8] Hosmer,D. Lemeshow,S. (1989). “*Applied Logistic Regression*”. John Wiley Sons, Ltd.

[9] Humbert P, Dupond JL. “Multiple Autoimmune syndromes”*Ann Med Interne (Paris)*, (1988) 139:159-168.

[10] Kleinbaum, David G. (1997). “*Survival analysis: a self-learning text*”. New York, Springer-Verlag.

[11] Klein,J. Moeschberger,M. (2003). “*Survival Analysis*”.

[12] Mas Oliva,J. (2004).“*Diagnóstico molecular en medicina*”. México, El Manual Moderno.

APÉNDICE A

Programas

```
********************************************************************** *
Este programa calcula las diferentes dicotomizaciones de los marcadores,
para la metodología del modelo de regresión de Cox estratificado
y la regresión logística condicional.
**********************************************************************

Para Cox
library(RODBC)
canal=odbcConnectExcel2007(file.choose())
sqlTables(canal)
x=sqlFetch(canal,'Cox Mas Dico 012 tras ')
odbcCloseAll()
d=t(x[,2:dim(x)[2]])
```

```
nombres=x[,1]
colnames(d)=nombres
x=d dicot=function(X)ifelse(X==0,0,1)
dicot=function(X)ifelse(X==0,1,0)
dicot=function(X)ifelse(X==1,0,1)
dicot=function(X)ifelse(X==1,1,0)
dicot=function(X)ifelse(X==2,0,1)
dicot=function(X)ifelse(X==2,1,0)
base=apply(x,2,dicot)
write.csv(base,file.choose())
```

_____

Para Condicional

```
library(RODBC)
canal=odbcConnectExcel2007(file.choose())
sqlTables(canal)
x=sqlFetch(canal,'Condicional MAS Dico 012 tras ')
odbcCloseAll()
d=t(x[,2:dim(x)[2]])
nombres=x[,1]
colnames(d)=nombres
x=d
dicot=function(X)ifelse(X==0,0,1)
dicot=function(X)ifelse(X==0,1,0)
dicot=function(X)ifelse(X==1,0,1)
dicot=function(X)ifelse(X==1,1,0)
dicot=function(X)ifelse(X==2,0,1)
```

```
dicot=function(X)ifelse(X==2,1,0)
base=apply(x,2,dicot)
write.csv(base,file.choose())
```

********** *

Este programa permite dar la estructura que necesita SAS de la base de datos
para aplicar la metodología del modelo de regresión de Cox estratificado
y la regresión logística condicional.

****

```
Para Cox
library(RODBC)
canal=odbcConnectExcel2007(file.choose())
sqlTables(canal)
x=sqlFetch(canal,'Listo tras')
odbcCloseAll()
d=t(x[,2:dim(x)[2]])
nombres=x[,1]
colnames(d)=nombres
x=d
dim(x)
nomb=as.matrix(colnames(x))
b=x[,1:6]
m=rep(0,1,8)
for(i in 7:dim(x)[2])
MARCADOR=rep(nomb[i,1],68)
DICOT=x[,i]
m=rbind(m,cbind(b,MARCADOR,DICOT))
```

```
granmarca=m[-1,]
granmarca=as.matrix(granmarca)
write.csv(granmarca,file.choose())
*_____*
```

Para Condicional

```
library(RODBC)
canal=odbcConnectExcel2007(file.choose())
sqlTables(canal)
x=sqlFetch(canal,'Listo tras')
odbcCloseAll()
d=t(x[,2:dim(x)[2]])
nombres=x[,1]
colnames(d)=nombres
x=d
dim(x)
nomb=as.matrix(colnames(x))
b=x[,1:6]
m=rep(0,1,8)
for(i in 7:dim(x)[2])
MARCADOR=rep(nomb[i,1],58)
DICOT=x[,i]
m=rbind(m,cbind(b,MARCADOR,DICOT))
granmarca=m[-1,]
write.csv(granmarca,file.choose())
********************************************************************** *
```

Este programa calcula la regresión logística condicional

y el modelo de regresión de Cox estratificado.

****

```
/* Regresión logística condicional*/
data base;
set base;
switch=2;
if cc=1 then switch=1;
ods html;
proc sort data=base;
by MARCADOR;
run;
proc phreg;
by MARCADOR;
model switch*CC(0)=DICOT sexo/ties=discrete rl;
strata FAM;
run;
ods html close;
*_____*

/* modelo de regresión de Cox estratificado*/
ods html;
proc sort data=base;
by MARCADOR;
run;
proc phreg data=base;
by MARCADOR;
model T*censura(1)=DICOT sexo/rl;
```

```
strata FAM;
run;
ods html close;
```

APÉNDICE B

Lista de Marcadores Genéticos

N°	**MARCADOR**	N°	**MARCADOR**	N°	**MARCADOR**
1	ATA79C10	132	GATA11A11P	263	GATA32F05
2	GATA29A01	133	AAT013	264	ATA29A06P
3	GATA41G07M	134	AAAT072	265	GATA23C03P
4	GATA7G10	135	UT6540	266	ATA5A09N
5	GGAA21G11L	136	TTA032z	267	GGAA29H03N
6	GATA71H05	137	ATTT030	268	GATA86H01
7	ATA7D07	138	ATA50C05ZP	269	GATA11C08P
8	GATA188F04	139	GATA163B10N	270	GATA64F08
9	GTTTT002P	140	ATA12D05P	271	GATA43H03N
10	GGAA3A07M	141	GATA61E03	272	ATA26D07
11	GATA27E01	142	GGAT3H10M	273	GATA51B02ZP

12	GATA29A05P	143	GATA11E02N	274	GGAA22G01ZP
13	ATA47D07	144	GATA64D02	275	AGAT113Z
14	300wb9	145	ATA28B11	276	GATA74E02Z
15	GATA129H04	146	GATA68H04	277	ATA77F05Z
16	GATA72H07	147	ATA11D10Z	278	GATA43H01M
17	GATA26G09P	148	GATA31	279	ATA29G03Z
18	GATA152F05L	149	GATA23F08	280	GATA90G11M
19	GATA109Z	150	GATA32B03	281	GGAA30H04ZP
20	GATA6A05	151	GATA184A08	282	ATA19H08
21	GATA124C08N	152	GATA165G02M	283	GGAA4A12
22	GATA133A08Q	153	ATA6C09P	284	GATA169E06ZP
23	ATA25E07M	154	GATA81B01	285	GATA193A07
24	GATA12A07N	155	ATA22G07P	286	GATA168F06
25	GATA43A04	156	035xb9ZP	287	ATGG002
26	GGAA5F09	157	GATA24F03ZP	288	ATT198Z
27	GGAA22G10N	158	GATA119B03	289	AATA036
28	TATC028	159	GATA137H02N	290	GATA143C02
29	ATA4E02	160	GGAA3F06	291	GATA88H02N
30	GATA7C01	161	GATA13G11ZP	292	GATA50C03N
31	GATA48B01	162	GATA31A10	293	GATA63A03N
32	GATA124F08	163	GATA24D12P	294	GATA50G06
33	GATA4H09	164	GATA118G10	295	GATA151F03N
34	ATA29C07L	165	GATA73D10L	296	GATA85D02
35	ATA009	166	GATA3F01	297	204ZG5ZP
36	GATA22D12	167	GATA5D08	298	ATA24A08

37	GATA50F11	168	GATA23F05	299	GATA73F01M
38	SraP	169	GGAA6D03N	300	GATA197B10P
39	130yg9P	170	GATA43C11	301	GATA27A03
40	GATA116B01N	171	GATA63F08P	302	TTTA028
41	GGAA20G10M	172	GATA104	303	ATA41E04
42	GATA11H10	173	GATA189C06M	304	ATA3A07
43	GATA8F07	174	GATA30D09N	305	TTAT023Z
44	GATA86E02P	175	MFD442-GTTT002	306	ATT001
45	ATA47C04P	176	ATT023	307	CATA002Z
46	ATA27D04P	177	TTCA004P	308	GGAA3G05
47	GATA66D01ZP	178	ATT070Z	309	GATA22F09P
48	GATA69E12M	179	ATAA018P	310	AAT107Z
49	GATA88G05	180	UT7129L	311	GATA81D12M
50	GATA176C01	181	GGAA20C10Z	312	MFD466-TTA001
51	GATA4E11	182	GATA8G10M	313	ATACC001
52	GATA27A12	183	GGAA8G07	314	GATA11C06N
53	GATA4D07	184	GATA41A01	315	GATA71F09
54	GGAA20G04	185	GATA14E09	316	044xg3
55	ATA27H09	186	GATA8B01	317	GTAT1A05
56	GATA71D01	187	GAAT1A4N	318	GATA158H04
57	GATA65C03M	188	GATA26E03M	319	GATA8C04
58	GATA52A04M	189	GATA6B02P	320	ATA78D02N
59	GATA30E06P	190	GATA21C12	321	GATA185H04N
60	GATA4G12	191	GATA50D10	322	GGAA9D03
61	GATA23D03ZP	192	UT721M	323	GATA25A04

62	GATA12H10	193	MFD455-AAT052	324	095TC5ZP
63	GATA23A02	194	aaaac001	325	AAT245
64	GATA178G09M	195	GATA62F03M	326	GATA49C09N
65	AGAT021	196	GATA187D09N	327	300xa5P
66	GATA22G12	197	AGAT142P	328	GATA28D11
67	MFD433-AGAT010	198	GATA87E02N	329	TTCA006M
68	GATA131D09	199	GATA5E06P	330	GATA178F11z
69	295yc9P	200	GATA7D12	331	ATA45G06
70	079YG5ZP	201	GATA89A11	332	AGAT060
71	GATA73D01	202	GATA21F05	333	ACT1A01
72	GATA27C08P	203	GATA81C04M	334	GATA11A06
73	GATA8B05M	204	ATA18A07M	335	GATA64H04
74	ATA10H11	205	GATA27Z	336	GATA13
75	GATA6F06	206	GATA48D07	337	GATA6D09
76	AGAT128	207	GATA64G07	338	ATA23G05
77	AAC023	208	ATA59H06Z	339	GATA7E12
78	GATA7F05	209	TTTTA002	340	ATA82B02N
79	GATA128C02M	210	GATA88F09	341	GATA177C03N
80	GATA84B12	211	ATCC001	342	GATA44F10P
81	GATA68F07	212	ATA31G11P	343	GATA21G05
82	ATA34G06	213	GATA84C01ZP	344	GATA23B01N
83	GATA4A10	214	GATA70E11	345	GATA66B04
84	GATA3C02ZP	215	GATA73E11	346	GGAA2A03
85	AAT071	216	ATA5A04N	347	GATA156F11
86	GATA3H01	217	ATA21A03Z	348	UT7544

87	GATA22F11NZ	218	ATA24F10	349	Mfd232
88	TTTA040	219	GATA121A08N	350	GATA29B01L
89	GATA6G12	220	GATA87G01	351	Mfd238
90	ATA22E01	221	GGAT1A4	352	GATA51D03
91	4PTEL04	222	GATA115E01N	353	GATA72E11
92	GATA22G05M	223	GGAA2f11N	354	GATA129B03N
93	ATT015	224	GATA64A09	355	GATA29F06z
94	GATA70E01	225	GATA71C09	356	GATA42A03
95	ATA27C07P	226	ATA29C03	357	GATA47F05
96	GATA72G09Z	227	ATA22D02	358	AAT269
97	ATA21F01	228	GGAA23C05N	359	GATA45B10N
98	GATA28F03	229	ATGT006Z	360	UT254
99	GATA24H01N	230	GGAA17G05P	361	UT1772
100	GATA10G07	231	ATA33B03Z	362	GGAA3C07
101	ATA2A03	232	GATA23F06L	363	GATA129D11N
102	GATA2F11	233	GATA48E02	364	ATA27F01
103	GATA62A12Z	234	ATA34E08N	365	UT1355z
104	TAGA006	235	GATA6B09P	366	GATA70B08
105	ATA26B08	236	ATA1B07	367	GATA198B05N
106	GATA11E09	237	GATA63F09	368	AGAT120
107	GATA107	238	ATA9B04N	369	ATTT019M
108	GATA8A05	239	GATA46A12	370	GATA21F03
109	GATA27G03	240	GATA90D07N	371	GATA11B12
110	GGAA19H07	241	GATA30G01	372	ATA37D06
111	GATA42H02P	242	GATA28D01M	373	UT7136

112	165zf8ZP	243	GATA71E06	374	TCTA015M
113	GATA5B02M	244	GATA23E06L	375	GATA52B03
114	ATA20G07M	245	GATA64D03	376	AGAT144
115	GATA145D10N	246	GATA117D01N	377	GATA175D03
116	GATA84E11	247	ATA27C11ZP	378	ATA28C05
117	GATA3E10	248	GATA4H03	379	GATA124E07
118	GATA134B03	249	GATA49D12N	380	GATA027
119	GATA7C06M	250	Mfd259	381	GATA69C12
120	GATA21D04	251	GATA6C01	382	GATA144D04
121	MFD601	252	ATA27A06P	383	GATA72E05M
122	GATA67D03	253	GATA91H06M	384	GATA31D10M
123	GATA138B05ZP	254	UT5029	385	GATA31F01P
124	GATA52A12	255	GATA73H09N	386	GATA172D05
125	GATA89G08z	256	GATA26D02M	387	GATA48H04
126	GATA3H06M	257	GATA63D12P	388	GATA165B12P
127	GATA68A03	258	GATA85A04M	389	ATCT003
128	GATA62A04	259	PAH	390	GATA31E08
129	GATA2H09	260	ATA25F09M	391	TATC043
130	ATA23A10M	261	GGAA22C05	392	224zg11
131	GATA6E05	262	GATA4H01	393	TTTA062

APÉNDICE C

Tablas Descriptivas Familias Seleccionadas

Tabla C.1: Descripción familias con MAS, usadas para aplicar la metodología modelo de regresión de Cox estratificado

	Identificador Familia	**Número de Personas**	**Número de Mujeres**
	2005	2	2
	6058	3	1
	9002	7	5
	9004	2	1
	9009	2	2
	9013	2	2
	9014	4	3
	9016	6	6
	9019	3	2
	9024	7	5
	9025	4	3
	9030	4	3
	9045	5	5
	9047	2	2
	9051	4	1
	9053	4	3
	9058	2	2
	9060	3	3
	9064	2	2
Total	**19 Familias**	**68**	**53**

Tabla C.2: Descripción familias con MAS, usadas para aplicar la metodología regresión logística condicional

	Identificador Familia	**Número de Personas**	**Número de Mujeres**
	2005	2	2
	6058	3	1
	9002	7	5
	9004	2	1
	9009	2	2
	9013	2	2
	9014	4	3
	9016	5	5
	9019	3	2
	9024	5	3
	9025	2	2
	9030	3	2
	9045	2	2
	9047	2	2
	9051	3	1
	9053	4	3
	9058	2	2
	9060	3	3
	9064	2	2
Total	**19 Familias**	**58**	**45**

Tabla C.3: Descripción familias con SEMIMAS, usadas para aplicar la metodología del modelo de regresión de Cox estratificado

	Identificador Familia	**Número de Personas**	**Número de Mujeres**
	1124	2	1
	1138	3	2
	1147	4	3
	1173	3	3
	2005	2	2
	2026	2	1
	2252	2	2
	2267	4	3
	6058	3	1
	9001	8	5
	9002	7	5
	9004	2	1
	9007	5	4
	9009	2	2
	9013	2	2
	9014	4	3
	9015	4	3
	9016	6	6
	9019	3	2
	9022	8	8
	9024	7	5
	9025	4	3
	9030	4	3
	9039	6	3
	9045	5	5
	9047	2	2
	9051	4	1
	9053	4	3
	9058	2	2
	9060	3	3
	9064	2	2
Total	**31 Familias**	**119**	**91**

Tabla C.4: Descripción familias con MAS, usadas para aplicar la metodología regresión logística condicional

	Identificador Familia	**Número de Personas**	**Número de Mujeres**
	1124	2	1
	1138	3	2
	1147	3	3
	1173	3	3
	2005	3	2
	2026	2	1
	2252	2	2
	2267	3	3
	6058	3	1
	9001	4	4
	9002	7	5
	9004	2	1
	9007	5	4
	9009	2	2
	9013	2	2
	9014	4	3
	9015	4	3
	9016	5	5
	9019	3	2
	9022	2	2
	9024	5	3
	9025	2	2
	9030	3	2
	9039	4	3
	9045	2	2
	9047	2	2
	9051	3	1
	9053	4	3
	9058	2	2
	9060	3	3
	9064	2	2
Total	**31 Familias**	**96**	**76**

APÉNDICE D

Consentimiento Informado

CONSENTIMIENTO INFORMADO
SINDROME AUTOINMUNE MULTIPLE

CORPORACIÓN PARA INVESTIGACIONES BIOLÓGICAS (CIB)
MEDELLÍN, COLOMBIA

Nosotros le invitamos a Usted a tomar parte en un estudio de investigación

Si Ud. Decide tomar parte de este estudio le extraeremos 20cc de sangre hoy. Estamos particularmente interesados en las sustancias producidas por su organismo que pueden ser importantes para explicar su enfermedad. La muestra se tomará para realizar el estudio. De esta muestra se realizarán únicamente los análisis correspondientes a verificar los genes (el código o huella dactilar de las células) con

el objeto de encontrar cuales de estos genes se presentan más frecuentemente en pacientes con enfermedades como la que usted padece y de la misma muestra se analizarán otras sustancias (moléculas inflamatorias) que se presentan en pacientes con esta enfermedad. No se realizarán otros tipos de análisis genéticos ni otros experimentos con la misma muestra.

La muestra de sangre necesaria para analizar en el laboratorio será obtenida de la vena de su brazo. Esta es la manera usual como se obtiene sangre para el análisis. Le puede dar un poco de dolor cuando la aguja entre en su brazo. En una de 10 personas queda una pequeña cantidad de sangre debajo de la piel, lo cual causará un moretón. Hay un pequeño riesgo (1/100) de que la vena se coagule por un tiempo corto. El riesgo de infección, o pérdida de mucha sangre es muy bajo (menos de 1/100).

Todo lo que aprendemos de usted durante la investigación será confidencial. Si publicamos los resultados del estudio en una revista o libro científico, no lo identificaremos a usted de ninguna manera.

No le garantizamos que su participación en el estudio lo beneficie a usted. Ud. No recibirá ninguna compensación por participar en este estudio. Usted no tendrá costos adicionales por su participación en este estudio. Su decisión para tomar parte en este estudio es voluntaria. Usted tiene libertad de decidir si no quiere participar en este estudio en cualquier momento. Si decide no participar, o parar en cualquier momento, esto no afectará su cuidado medico futuro

Si tiene preguntas ahora, tiene la libertad de hacerlas. Si tiene preguntas adicionales mas tarde sobre la investigación que se hará en sus muestras, puede llamar al Dr. Juan Manuel Anaya al teléfono 4410855 Extensión 217 o 233. el comité de Ética de la Corporación para Investigaciones biológicas, CIB, que revisan los programas de investigación en humanos le responderá cualquier pregunta sobre sus derechos como

sujeto en esta investigación.

SU FIRMA INDICA QUE USTED DECIDIÓ TOMAR PARTE EN ESTA INVESTIGACIÓN Y QUE USTED HA LEIDO Y ENTENDIDO LA INFORMACIÓN AQUÍ SUMINISTRADA Y HA SIDO EXPLICADA A UD.

_____ _____- _____

Nombre Firma Cédula

_____ _____- _____

Testigo 1 (Acudiente) Firma Cédula

_____ _____- _____

Testigo 2 Firma Cédula

_____ _____- _____

Fecha Firma del Investigador Cédula