UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

MODELOS DE REGRESIÓN LOGÍSTICA APLICADOS AL ESTUDIO DE TRANSMISIÓN GENÉTICA

T E S I N A

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A:

SERGIO GUILLERMO SAAVEDRA FRANCO

DIRECTORA:
DRA. ELIANE REGINA RODRIGUES

CIUDAD UNIVERSITARIA, CDMX, 2017

“Son muchos los portentos, pero ninguno es superior al hombre”

Agradecimientos

A mi familia, por ser mi guía, mi apoyo y mi fuerza; gracias a ustedes tengo las herramientas que me permiten emprender, luchar y conseguir cualquier sueño. A mi madre, por tener el brillo más puro y radiante, por siempre procurarme con su inmenso amor y cariño, por darme siempre ánimos y por estar a mi lado en todo momento. A mi héroe, mi padre, por tus consejos, tu entrega, cariño y por estar siempre en pie de lucha por nosotros. A mi cómplice, Alan, por la alegría de tenerte en mi vida desde que me dieron la noticia de que tendría un hermano y por todas las aventuras que nos han hecho crecer y llegar a donde estamos.

A mi Cecilia, la mujer que cada día me hace sonreír y vivir, gracias por estar conmigo, por enseñarme junto a tus padres que con lucha, determinación y esfuerzo se obtiene cualquier meta, por creer en mi y darme la mano, "porque eres mi amor, mi cómplice y todo… y en la calle, codo a codo… somos mucho más que dos".

A todos mis amigos, por su cariño y buenos momentos.

A la Dra. Eliane por darme la oportunidad de llevar a cabo este proyecto, por sus invaluables enseñanzas, su paciencia y todos sus consejos.

A mis sinodales: Dra. María del Pilar Alonso Reyes, Dra. María Asunción Begoña Fernández Fernández, Dra. Ana Meda Guardiola y Act. Jaime Vázquez Alamilla por tener la disposición de apoyarme a mejorar esta tesina y por ser, mediante sus enseñanzas, un ejemplo a seguir para los estudiantes de la Facultad.

A la UNAM, en especial a la E.N.P. 3 "Justo Sierra" y a la Facultad de Ciencias, por todo lo que han significado en mi vida, por la excelente formación que he recibido, por las amistades encontradas y por el orgullo de ser universitario.

A Dios, por su amor infinito y por estar a mi lado.

Índice

Introducción 5
Objetivo 6
1 Conceptos básicos 7
1.1 Conceptos básicos de genética 7
1.2 Conceptos básicos de inferencia estadística 8
2 Resultados básicos 9
2.1 Modelos genéticos 9
2.2 Funciones de verosimilitud para pedigríes 11
2.3 Matrices de transmisión y transición genéticas 21
3 Modelos de regresión logística 25
3.1 Definición y resultados preliminares 25
3.2 Modelos de regresión logística 26
3.3 Estimación de parámetros de un modelo de regresión logística 33
4 Modelos de regresión logística aplicados a genética 37
4.1 Consideraciones preliminares 37
4.2 Modelo general 37
4.3 Modelos específicos 40
4.4 Modelos de regresión logística aplicados al análisis genético 44
Conclusiones 50
Apéndice 52
Bibliografía 53

Introducción

Siempre se ha tomado como cierto que la descendencia de los seres vivos hereda características de la especie. En 1865, Gregor Mendel brinda los principios básicos sobre los mecanismos de transmisión de rasgos en generaciones llevando a cabo experimentos, con rigor científico, sobre la reproducción de la planta de guisantes.

A partir de la investigación de Mendel, el campo de estudio ha crecido y obtenido sustanciosos resultados que han permitido avanzar en el diagnóstico y el tratamiento de las enfermedades hereditarias, mejorar la producción de alimentos y elaborar medicamentos mediante ingeniería genética. Por lo anterior, es de gran interés conocer las variables que intervienen en la transmisión de enfermedades en las generaciones, para así ubicar qué se puede hacer al respecto y, con ello, tratar el riesgo de manifestar patologías en la sociedad teniendo la finalidad de mejorar la salud pública.

La tesina se integra por cuatro capítulos. En el primero, se expone el universal conceptual usado en el documento, se señalan los referentes básicos usados para la comprensión del tema tratado.

En el segundo, se diserta sobre cómo se heredan de generación en generación los rasgos genéticos mediante la descripción de los modelos de transmisión genética.

La tercera sección desarrolla los modelos de regresión logística, que es la metodología estadística con la cual se modelará la relación entre los factores de riesgo y la presencia o no de rasgos genéticos, se expone sobre la razón de momios, así como de la estimación y comprobación de dichos modelos.

Finalmente, el cuarto capítulo vincula al modelo de regresión logística con los modelos de transmisión genética. Se construyen distintos tipos de modelos que permiten conocer qué factores de riesgo, entre los que destaca la herencia biológica, tienen algún grado de influencia en manifestar o no algún rasgo o enfermedad en estudio. Se indica el uso adecuado para cada tipo de modelo y se muestra un ejemplo de la metodología expuesta.

Objetivo

Hoy en día, la estadística desempeña un rol de suma significancia en diversas áreas del conocimiento. Ella es necesaria para obtener conclusiones de eventos de interés mediante la formulación de modelos que expliquen, de la mejor forma posible, el comportamiento de lo estudiado.

A su vez, la genética es una ciencia en constante crecimiento, con recientes hallazgos y muchos cuestionamientos por resolver. Por ello, el empleo de herramientas estadísticas es fundamental para lograr el completo entendimiento de los mecanismos de transmisión en las especies. Por ende, la motivación del presente proyecto, es mostrar un ejemplo del uso de la estadística en el estudio de algunos problemas en genética, particularmente en la comprensión del funcionamiento de la transmisión genética de enfermedades mediante el uso de modelos de regresión logística.

1 Conceptos básicos

1.1 Conceptos básicos de genética

El objetivo principal de la **genética** es estudiar la forma en que las características de los organismos vivos se transmiten y se expresan, de una generación a otra, bajo diferentes condiciones ambientales. De esta forma, la genética busca comprender y explicar la herencia biológica existente en las distintas generaciones de individuos y, con esto, describir la respuesta a los cuestionamientos del tipo: ¿qué se hereda? y ¿cómo se hereda?

Para comprender los mecanismos por los que se transmiten rasgos, Pierce (2009) define los siguientes conceptos:

La **herencia** es el conjunto de caracteres que transmiten los padres a los hijos. Los **cromosomas** son los vehículos de información genética dentro de la célula. Las unidades fundamentales de la herencia son los **genes,** los cuales se encargan de codificar las características genéticas**.** El **locus** es el lugar específico ocupado por un gen en el cromosoma. Las formas alternativas de un mismo gen se denominan **alelos**. Por ejemplo, un gen codifica el color de la piel de forma general y los alelos sus distintas variaciones (colores).

Hay distinción entre los **rasgos** y los **genes**. La expresión de los rasgos en un individuo depende de los genes heredados y factores ambientales. Entonces, el conjunto de alelos que posee un organismo, la información genética heredada, es el **genotipo;** mientras que la apariencia o manifestación de una característica es el **fenotipo.**

Por ejemplo, el grupo sanguíneo A es un fenotipo mientras que la información genética que codifica el grupo sanguíneo es el genotipo. Es decir, cada fenotipo es el resultado de un genotipo desarrollado en un ambiente específico. Se hereda el genotipo, pero no el fenotipo.

Cada individuo hereda de sus padres un alelo para cada tipo de gen en su correspondiente locus en un cromosoma.

1.2 Conceptos básicos de inferencia estadística

La estadística inferencial es una parte de la Estadística que comprende los métodos y procedimientos necesarios para hacer inferencias (deducir propiedades) de una población a partir de una muestra de la misma. Pretende crear modelos que permitan entender diversos comportamientos sobre las variables estudiadas. La evaluación de qué tan buenas son las deducciones obtenidas es medida en términos probabilísticos; las inferencias vienen de la mano con su probabilidad de acierto.

La función de verosimilitud es una función de los parámetros de un modelo estadístico y permite realizar inferencias acerca de su valor a partir de un conjunto de observaciones. El método de máxima verosimilitud es de gran uso en la obtención de los estimadores correctos de modelos estadísticos. A partir de la función de verosimilitud es posible obtener los valores de los estimadores de los parámetros del modelo considerando que son los que maximizan la probabilidad de que la muestra estudiada es obtenida de un fenómeno que se puede describir usando el modelo planteado.

Para hacer un buen uso del modelado estadístico, se deben conocer los supuestos mediante los cuales el modelo elegido tendrá un buen ajuste respecto de los datos en estudio.

De esta forma, si $x_1, x_2, \dots, x_n$ es una muestra obtenida de observar un fenómeno descrito por un modelo, con vector de parámetros θ, entonces si $x_1, x_2, \dots, x_n$ son independientes, la función de verosimilitud es

$$f(x_1, x_2, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

Si la muestra no es independiente, entonces se deben realizar ajustes de acuerdo con las dependencias existentes entre observaciones.

La correcta implementación de un modelo hace que se pueda inferir sobre los parámetros, permitiendo conocer las variables explicativas que afectan a la respuesta y la magnitud en que lo hacen.

2 Resultados básicos

2.1 Modelos genéticos

El uso de herramientas estadísticas para realizar estudios en el ámbito de la genética es de gran importancia para entender y atender enfermedades hereditarias. Se presentan a continuación modelos que explican la manera en que se heredan los rasgos y como calcular la incidencia de éstos en pedigríes.

Con el término **pedigrí**, se hará referencia al árbol genealógico de un individuo a partir de una pareja de fundadores. Un pedigrí se compone de dos tipos de miembros, los que son descendientes de cualquier apareamiento de la generación anterior y los que se convierten en parejas de estos descendientes. En el presente estudio no se toman en cuenta matrimonios consanguíneos.

Se tomará la notación de Elston y Stewart (1971). Se denota con $\boldsymbol{X}$ a los individuos que tienen progenitores en el pedigrí y con $\boldsymbol{Y}$ a quienes entran en el pedigrí por ser parejas de individuos denotados por $\boldsymbol{X}$ ya pertenecientes. Los progenitores de toda la descendencia, la pareja original o fundadora, serán X_i y Y_i.

Se hará uso de subíndices para representar de forma precisa a cada integrante. El subíndice asignado muestra la generación a la que pertenece cada miembro.

En la figura 2.1 están algunos ejemplos de pedigríes.

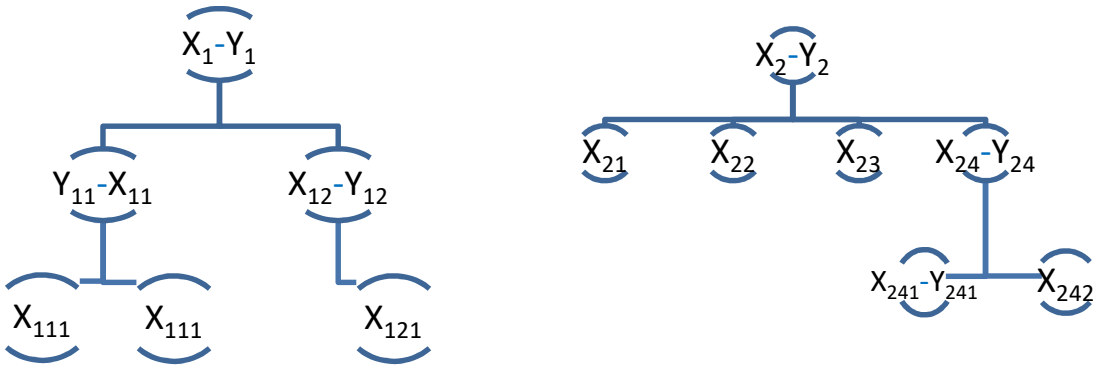

Figura 2.1 Ejemplos de pedigrís. Elston y Stewart (1971).

La posición del número en el subíndice representa la generación del individuo. Por lo tanto, si sólo hay un dígito, entonces es la generación concerniente a los progenitores fundadores X_1 y Y_1. Si hay dos dígitos, se habla de la primera generación, correspondiente a los descendientes de la pareja original y, si las hay, sus respectivas parejas, es decir, $(x_{11,}x_{12},\dots,x_{1n},y_{11,}y_{12},\dots,y_{1n})$. Al tener tres dígitos se hace referencia a la segunda generación, siendo los nietos de los progenitores originales y sus parejas $(x_{1\,i_1\,i_2}\ ,\ y_{1\,i_1\,i_2})$. Con lo cual x_{123}, es el tercer hijo del segundo hijo de los padres fundadores.

Por lo tanto, la notación para representar a un individuo perteneciente a la n-ésima generación de un pedigrí es:

$X_{i_0\,i_1\,i_2\dots i_n}$ si sus progenitores pertenecen al pedigrí

$Y_{i_0\,i_1\,i_2\dots i_n}$ si es pareja de alguno de los individuos con progenitores en el pedigrí.

2.1.1 Modelo básico

Cannings et al. (1978) diserta sobre un modelo básico para la transmisión genética de individuo a individuo. A continuación se describen los supuestos y la notación básica del modelo:

(i) Cada individuo A miembro del pedigrí se representa por la tripleta:

$$(\sigma(\mathrm{A}),\ \varphi(\mathrm{A}),\ \mathrm{e}(\mathrm{A})),$$

de la cual, $\sigma(\mathrm{A})$ especifica su genotipo; $\varphi(\mathrm{A})$ su fenotipo y $\mathrm{e}(\mathrm{A})$ sus características ambientales como edad, sexo, información sobre consumo o exposición a sustancias, entre otras.

(ii) Para cada individuo, el fenotipo depende del genotipo y de las variables ambientales. De esta forma, se definen las **probabilidades de penetrancia**:

$$\Pr(x \mid g, \mathrm{e}(\mathrm{A})) = g_g(\mathrm{x}) = \Pr(\varphi(\mathrm{A}) = x \mid \sigma(\mathrm{A}) = \mathrm{g}, \mathrm{e}(\mathrm{A})),$$

es decir, la probabilidad de que el individuo presente el fenotipo x dadas las variables ambientales y su genotipo g.

Y si no se observa el fenotipo x de A, entonces

$$g_g(\mathrm{x}) = 1 \text{ para todo genotipo } g.$$

(iii) Dado el genotipo de A, su fenotipo es independiente de los fenotipos de cualesquiera otros individuos.

(iv) El genotipo de A solo depende del genotipo de su padre σ(P) y el de su madre σ(M). Se define la **probabilidad de transmisión**:

$$P_{jki} = \Pr(\sigma(\mathrm{A}) = g \mid \sigma(\mathrm{M}) = j, \sigma(\mathrm{P}) = \mathrm{k}),$$

es decir, la probabilidad de que un individuo tenga genotipo g dado que sus progenitores tienen genotipos j y k.

(v) Dados los genotipos de los progenitores, los genotipos de su descendencia son mutuamente independientes.

A partir del modelo básico, Cannings et al. (1978) plantea modelos genéticos dentro de los cuales se encuentran los dos siguientes:

Modelo de un locus, dos alelos

A partir de los dos alelos **A**, **a** se tienen tres genotipos, siendo éstos (**AA, Aa, aa**).

Modelo de un locus, n-alelos

A partir de n alelos $a_1, a_2, \ldots, a_n$, hay $\frac{1}{2}n(n+1)$ genotipos que son las combinaciones con repetición de n genotipos tomados dos a dos.

2.2 Funciones de verosimilitud para pedigríes

Suponga que existen k genotipos posibles que se enumeran 1,2,…,k. Para construir la verosimilitud de observar un fenotipo dado en el i-ésimo hijo, se supone que cada individuo tiene uno de entre los k genotipos causantes de variación en el rasgo estudiado.

Al igual que en la sección 2.1, se denota como P_{stu} a la probabilidad de que el descendiente tenga genotipo u dado que los genotipos de sus progenitores son s y t, es decir,

$$P_{stu} = \Pr(u \mid s, t).$$

También como en la sección 2.1, se toma $g_u(x_i)$ como la probabilidad de que el i-ésimo hijo presente el fenotipo x_i dado que tiene genotipo u, es decir,

$$g_u(x_i) = \Pr(\varphi(X_i) = x_i \mid u)$$

Para los siguientes desarrollos, se tomarán en cuenta las siguientes justificaciones:

(i) Se usa la ley de probabilidad total.

(ii) Se utilizan propiedades de probabilidad condicional, es decir, $\Pr(A \cap B) = \Pr(A|B)\Pr(B)$.

(iii) Se toma en cuenta la independencia de eventos de acuerdo a los supuestos del modelo, en particular, el supuesto (v) de la sección 2.1.

(iv) Utilizando el hecho de que el fenotipo de un individuo solamente depende de su genotipo.

(v) Reacomodo de la expresión.

Con lo anterior, la probabilidad de que un individuo X_i tenga fenotipo x_i dado que los genotipos de sus padres son s y t es:

$$\begin{aligned}
&\Pr(\varphi(X_i) = x_i \mid \sigma(M_{X_i}) = s\ , \sigma(P_{X_i}) = t) \\
&= \Pr(x_i \mid s,t) \\
\text{(i)} \quad &= \sum_{u=1}^{k} \Pr(x_i, u \mid s,t) \\
\text{(ii)} \quad &= \sum_{u=1}^{k} \Pr(x_i \mid u,s,t)\Pr(u \mid s,t) \\
\text{(iv)} \quad &= \sum_{u=1}^{k} \Pr(x_i \mid u)\Pr(u \mid s,t) \\
&= \sum_{u=1}^{k} g_u(x_i)\ P_{stu} \\
\text{(v)} \quad &= \sum_{u=1}^{k} P_{stu}\ g_u(x_i).
\end{aligned} \tag{2.1}$$

Conociendo lo anterior y con la independencia entre los genotipos de los descendientes dados los genotipos de los progenitores citada en el modelo básico, la función de verosimilitud conjunta de observar los rasgos $x_1, x_2, \dots, x_n$ en la descendencia $X_1, X_2, \dots, X_n$ dado que se conocen los genotipos de sus progenitores es:

$$\Pr(\varphi(X_1) = x_1, \varphi(X_2) = x_2, \dots, \varphi(X_n) = x_n \,|\, \sigma(M_X) = s\ , \sigma(P_X) = t))$$

$$= \Pr(\varphi(X_1) = x_1, \varphi(X_2) = x_2, \dots, \varphi(X_n) = x_n\ \,|\, s\ , t)$$

$$= \Pr(x_1, x_2, \dots, x_n \,|\, s\ , t)$$

(i) $$= \sum_{u_1=1}^{k} \sum_{u_2=1}^{k} \dots \sum_{u_n=1}^{k} \Pr(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_n \,|\, s\ , t)$$

(iii) $$= \sum_{u_1=1}^{k} \sum_{u_2=1}^{k} \dots \sum_{u_n=1}^{k} \Pr(x_1, u_1|s, t) \Pr(x_2, u_2|s, t) \dots \Pr(x_n, u_n|s, t)$$

(v) $$= \sum_{u_1=1}^{k} \Pr(x_1, u_1|s, t) \sum_{u_2=1}^{k} \Pr(x_2, u_2|s, t) \dots \sum_{u_n=1}^{k} \Pr(x_n, u_n|s, t)$$

$$= \sum_{u=1}^{k} \Pr(x_1, u|s, t) \sum_{u=1}^{k} \Pr(x_2, u|s, t) \dots \sum_{u=1}^{k} \Pr(x_n, u|s, t)$$

$$= \prod_{i=1}^{n} \left[\sum_{u=1}^{k} \Pr(x_i, u|s, t) \right]$$

(ii) $$= \prod_{i=1}^{n} \left[\sum_{u=1}^{k} \Pr(x_i|u, s, t) \Pr(u|s, t) \right]$$

(iv) $$= \prod_{i=1}^{n} \left[\sum_{u=1}^{k} \Pr(x_i|u) \Pr(u|s, t) \right]$$

$$= \prod_{i=1}^{n} \left[\sum_{u=1}^{k} g_u(x_i)\, P_{stu} \right]$$

$$\overset{(v)}{=} \prod_{i=1}^{n} \left[\sum_{u=1}^{k} P_{stu} \;\; g_u(x_i) \right]. \tag{2.2}$$

Se tiene que $g_v(y_i)$ es la probabilidad del individuo i de presentar el fenotipo y_i dado el genotipo v. Se define a Ψ_v como la probabilidad de que un individuo en la población tenga genotipo v. De esta forma, se obtiene la función de verosimilitud de observar la característica en la pareja del i-ésimo miembro de la descendencia:

$$\begin{aligned}
\Pr(\varphi(Y_i) = y_i) \;\; &= \Pr(y_i) \\
&\overset{(i)}{=} \sum_{v=1}^{k} \Pr(y_i, v) \\
&\overset{(ii)}{=} \sum_{v=1}^{k} \Pr(y_i|v) \;\Pr(\mathrm{v}) \\
&= \sum_{v=1}^{k} \;\; g_v(y_i)\, \Psi_v \\
&\overset{(v)}{=} \sum_{v=1}^{k} \Psi_v \;\; g_v(y_i).
\end{aligned} \tag{2.3}$$

Así, encontramos la función de verosimilitud de las n parejas de una descendencia, tomando en cuenta la independencia existente en los genotipos de individuos sin parentesco entre sí.

$$\begin{aligned}
&\Pr(\varphi(Y_1) = y_1, \varphi(Y_2) = y_2, \dots, \varphi(Y_n) = y_n) \\
&= \Pr(y_1, y_2, \dots, y_n)
\end{aligned}$$

$$
\begin{aligned}
\text{(i)} \quad &= \sum_{v_1=1}^{k} \sum_{v_2=1}^{k} \dots \sum_{v_n=1}^{k} \Pr(y_1, y_2, \dots, y_n, v_1, v_2, \dots, v_n) \\
\text{(iii)} \quad &= \sum_{v_1=1}^{k} \sum_{v_2=1}^{k} \dots \sum_{v_n=1}^{k} \Pr(y_1, v_1) \Pr(y_2, v_2) \dots \Pr(y_n, v_n) \\
\text{(v)} \quad &= \sum_{v_1=1}^{k} \Pr(y_1, v_1) \sum_{v_2=1}^{k} \Pr(y_2, v_2) \dots \sum_{v_n=1}^{k} \Pr(y_n, v_n) \\
&= \sum_{v=1}^{k} \Pr(y_1, v) \sum_{v=1}^{k} \Pr(y_2, v) \dots \sum_{v=1}^{k} \Pr(y_n, v) \\
&= \prod_{i=1}^{n} \left[\sum_{v=1}^{k} \Pr(y_i, v) \right] \\
\text{(ii)} \quad &= \prod_{i=1}^{n} \left[\sum_{v=1}^{k} \Pr(y_i | v) \Pr(v) \right] \\
&= \prod_{i=1}^{n} \left[\sum_{v=1}^{k} g_v(y_i) \ \Psi_v \right] \\
\text{(v)} \quad &= \prod_{i=1}^{n} \left[\sum_{v=1}^{k} \Psi_v \ g_v(y_i) \right].
\end{aligned}
\tag{2.4}
$$

Con lo anterior, se obtiene la función de verosimilitud de observar el fenotipo en los hermanos en una descendencia y sus parejas dado que los fundadores del pedigrí tienen genotipos s y t.

$$
\begin{aligned}
&\Pr\left(\varphi(X_1) = x_1, \varphi(X_2) = x_2, \dots, \varphi(X_n) = x_n, \varphi(Y_1) = y_1, \right. \\
&\quad \left. \varphi(Y_2) = y_2, \dots, \varphi(Y_n) = y_n \mid \sigma(M_X) = s \ , \sigma(P_X = t)\right) \\
&= \Pr(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n | s, t)
\end{aligned}
$$

$$= \sum_{u_1=1}^{k} \sum_{u_2=1}^{k} ... \sum_{u_n=1}^{k} \sum_{v_1=1}^{k} \sum_{v_2=1}^{k} ... \sum_{v_n=1}^{k} \Pr(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n, u_1 \\ u_2, ..., u_n, v_1, v_2, ..., v_n | s, t)$$

$$= \sum_{u_1=1}^{k} \sum_{u_2=1}^{k} ... \sum_{u_n=1}^{k} \sum_{v_1=1}^{k} \sum_{v_2=1}^{k} ... \sum_{v_n=1}^{k} \Pr(x_1, u_1, x_2, u_2, ..., x_n, u_n, y_1, \\ v_1, y_2, v_2, ..., y_n, v_n | s, t)$$

$$= \sum_{u_1=1}^{k} \Pr(x_1, u_1|s,t) ... \sum_{u_n=1}^{k} \Pr(x_n, u_n|s,t) \sum_{v_1=1}^{k} \Pr(y_1, v_1|s,t) ...$$

$$\sum_{v_n=1}^{k} \Pr(y_n, v_n|s,t)$$

$$= \sum_{u=1}^{k} \Pr(x_1, u|s,t) ... \sum_{u=1}^{k} \Pr(x_n, u|s,t) \sum_{v=1}^{k} \Pr(y_1, v|s,t) ... \sum_{v=1}^{k} \Pr(y_n, v|s,t)$$

$$= \prod_{i=1}^{n} \left\{ \left[\sum_{u=1}^{k} \Pr(x_i, u|s,t) \right] \left[\sum_{v=1}^{k} \Pr(y_i, v|s,t) \right] \right\}$$

$$= \prod_{i=1}^{n} \left\{ \left[\sum_{u=1}^{k} \Pr(x_i|u)\ \Pr(u|s,t) \right] \left[\sum_{v=1}^{k} \Pr(y_i|v) \Pr(v|s,t) \right] \right\}$$

se tiene la igualdad $\Pr(v|s,t) = \Pr(v)$ debido a que los genotipos s y t de los padres de la hermandad integrada por los X_i no tienen influencia en la probabilidad de que las parejas Y_i presenten el genotipo v.

$$= \prod_{i=1}^{n} \left\{ \left[\sum_{u=1}^{k} g_u(x_i)\ P_{stu} \right] \left[\sum_{v=1}^{k} g_v(y_i)\ \Pr(v) \right] \right\}$$

$$= \prod_{i=1}^{n} \left\{ \left[\sum_{u=1}^{k} g_u(x_i)\ P_{stu} \right] \left[\sum_{v=1}^{k} g_v(y_i)\ \Psi_v \right] \right\}$$

$$= \prod_{i=1}^{n} \left\{ \left[\sum_{u=1}^{k} P_{stu}\ g_u(x_i) \right] \left[\sum_{v=1}^{k} \Psi_v\ g_v(y_i)\ \right] \right\}.$$

De forma general, renombrando los valores con la notación establecida en la sección 2.1 y situándonos en la j-ésima generación, se tiene la verosimilitud de observar el fenotipo en los individuos de la j-ésima generación del pedigrí dados los genotipos de sus progenitores, dada por:

$$\Pr\Big(x_{i_0 i_1 \ldots i_j=1}, \ldots, x_{i_0 i_1 \ldots i_j=n}, y_{i_0 i_1 \ldots i_j=1}, \ldots, y_{i_0 i_1 \ldots i_j=1} \,\Big|\, s_{j-1}\ t_{j-1}\Big)$$

$$= \prod_{i_j=1}^{n} \quad \sum_{s_j=1}^{k} P_{s_{j-1}\ t_{j-1}\ s_j} \ \ g_{s_j}(x_{i_0 i_1 \ldots i_j}) \ \ \sum_{t_j=1}^{k} \Psi_{t_j}\ g_{t_j}(y_{i_0 i_1 \ldots i_j}).$$

Ejemplo 2.1:

Se procederá a calcular la probabilidad de que todos los miembros del pedigrí a la derecha en la figura 2.1 manifiesten fenotipos x_2, y_2, x_{21}, x_{22}, x_{23}, $x_{24}, y_{24}, x_{241}, y_{241}, x_{242}$ de acuerdo con las justificaciones dadas en el principio de esta sección y los supuestos establecidos en la sección 2.1. La probabilidad por obtener es:

$$\Pr(x_2, y_2, x_{21}, x_{22}, x_{23}, x_{24}, y_{24}, x_{241}, y_{241}, x_{242}) =$$

$$\overset{(i)}{=} \sum_{s_0=1}^{k} \sum_{t_0=1}^{k} \Pr(\, x_2, y_2, x_{21}, x_{22}, x_{23}, x_{24}, y_{24}, x_{241}, y_{241}, x_{242}, s_0, t_0)$$

$$\overset{(ii)}{=} \sum_{s_0=1}^{k} \sum_{t_0=1}^{k} \Pr(\, x_2 \mid y_2, x_{21}, x_{22}, x_{23}, x_{24}, y_{24}, x_{241}, y_{241}, x_{242}, s_0, t_0) *$$
$$\Pr(y_2, x_{21}, x_{22}, x_{23}, x_{24}, y_{24}, x_{241}, y_{241}, x_{242}, s_0, t_0)$$

$$\overset{(iv)}{=} \sum_{s_0=1}^{k} \sum_{t_0=1}^{k} \Pr(\, x_2 | s_0) \Pr(y_2, x_{21}, x_{22}, x_{23}, x_{24}, y_{24}, x_{241}, y_{241}, x_{242}, s_0, t_0)$$

De forma análoga, se obtiene la probabilidad de que Y_2 tenga fenotipo y_2 dependiendo únicamente de su genotipo:

$$= \sum_{s_0=1}^{k}\sum_{t_0=1}^{k} \Pr(\, x_2|s_0)\ \Pr(y_2|t_0)\Pr(x_{21},x_{22},x_{23},x_{24},y_{24},x_{241},y_{241},x_{242},s_0,t_0)$$

$$= \sum_{s_0=1}^{k}\sum_{t_0=1}^{k} g_{s_0}(x_2)g_{t_0}(y_2)\Pr(x_{21},x_{22},x_{23},x_{24},y_{24},x_{241},y_{241},x_{242},s_0,t_0)$$

$$= \sum_{s_0=1}^{k} g_{s_0}(x_2)\sum_{t_0=1}^{k} g_{t_0}(y_2)\Pr(x_{21},x_{22},x_{23},x_{24},y_{24},x_{241},y_{241},x_{242},s_0,t_0).$$

(2.5)

Se obtiene de (2.5) que

$$\Pr(x_{21},x_{22},x_{23},x_{24},y_{24},x_{241},y_{241},x_{242},s_0,t_0)$$

$$\overset{(i)}{=} \sum_{u_{21}=1}^{k}\sum_{u_{22}=1}^{k}\sum_{u_{23}=1}^{k}\sum_{u_{24}=1}^{k}\Pr\big(x_{21},u_{21},x_{22},u_{22},x_{23},u_{23},x_{24},u_{24},y_{24},x_{241},y_{241},x_{242},s_0,t_0\big)$$

$$\underset{(v)}{\overset{(ii,iv)}{=}} \sum_{u_{21}=1}^{k}\Pr(x_{21}|u_{21})\sum_{u_{22}=1}^{k}\Pr(x_{22}|u_{22})\sum_{u_{23}=1}^{k}\Pr(x_{23}|u_{23})\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\ *$$
$$\Pr(u_{21},u_{22},u_{23},u_{24},y_{24},x_{241},y_{241},x_{242},s_0,t_0)$$

$$= \sum_{u_{21}=1}^{k}\Pr(x_{21}|u_{21})\Pr(u_{21}|s_0,t_0)\sum_{u_{22}=1}^{k}\Pr(x_{22}|u_{22})\Pr(u_{22}|s_0,t_0)\ *$$
$$\sum_{u_{23}=1}^{k}\Pr(x_{23}|u_{23})\Pr(u_{23}|s_0,t_0)\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\Pr\big(y_{24},x_{241},y_{241},x_{242},s_0,t_0,u_{24}\big)$$

$$= \sum_{u_{21}=1}^{k} g_{u_{21}}(x_{21})P_{s_0t_0u_{21}}\sum_{u_{22}=1}^{k} g_{u_{22}}(x_{22})P_{s_0t_0u_{22}}\sum_{u_{23}=1}^{k} g_{u_{23}}(x_{23})P_{s_0t_0u_{23}}$$
$$\sum_{u_{24}=1}^{k} g_{u_{24}}(x_{24})\Pr(y_{24},x_{241},y_{241},x_{242},s_0,t_0,u_{24})$$

$$= \Big[\prod_{i=1}^{3}\Big(\sum_{u=1}^{k} g_{u_{2j}}(x_{2i})P_{s_0t_0u_{2j}}\Big)\Big]\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\sum_{v_{24}=1}^{k}\Pr(y_{24},v_{24},x_{241},y_{241},x_{242},s_0,t_0,u_{24})$$

$$= \Big[\prod_{i=1}^{3}\Big(\sum_{u=1}^{k} g_{u}(x_{2i})P_{s_0t_0u}\Big)\Big]\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\sum_{v_{24}=1}^{k}\Pr(y_{24},v_{24},x_{241},y_{241},x_{242},s_0,t_0,u_{24})$$

$$= \left[\prod_{i=1}^{3}\left(\sum_{u=1}^{k} g_u(x_{2i})P_{s_0t_0u}\right)\right]\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\sum_{v_{24}=1}^{k}\Pr(y_{24}|v_{24})\ \Pr(v_{24},x_{241},y_{241},x_{242},s_0,t_0,u_{24})$$

$$= \left[\prod_{i=1}^{3}\left(\sum_{u=1}^{k} g_u(x_{2i})P_{s_0t_0u}\right)\right]\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\sum_{v_{24}=1}^{k} g_{v_{24}}(y_{24})\Pr(x_{241},y_{241},x_{242},s_0,t_0,u_{24},v_{24}) \tag{2.6}$$

Se obtiene de (2.6) que

$$\Pr(x_{241},y_{241},x_{242}\,,s_0,t_0,u_{24},v_{24}) =$$

$$\overset{(i)}{=}\sum_{u_{241}=1}^{k}\sum_{u_{242}=1}^{k}\Pr(x_{241},u_{241},y_{241},x_{242}\,,u_{242},s_0,t_0,u_{24},v_{24})$$

$$\overset{(ii,iv)}{=}\sum_{u_{241}=1}^{k}\sum_{u_{242}=1}^{k}\Pr(x_{241}|u_{241})\Pr(x_{242}|u_{242})\Pr(u_{241},y_{241}\,,u_{242},s_0,t_0,u_{24},v_{24})$$

$$=\sum_{u_{241}=1}^{k}\sum_{u_{242}=1}^{k}\Pr(x_{241}|u_{241})\Pr(x_{242}|u_{242})\Pr(u_{241}|u_{24},v_{24})\Pr(u_{242}|u_{24},v_{24})\Pr(y_{241}\,,u_{242},s_0,t_0,u_{24},v_{24})$$

$$=\sum_{u_{241}=1}^{k}\sum_{u_{242}=1}^{k}\Pr(x_{241}|u_{241})\Pr(u_{241}|u_{24},v_{24})\Pr(x_{242}|u_{242})\Pr(u_{242}|u_{24},v_{24})\Pr(y_{241}\,,s_0,t_0,u_{24},v_{24})$$

$$=\sum_{u_{241}=1}^{k}\sum_{u_{242}=1}^{k} g_{u_{241}}(x_{241})P_{u_{24}v_{24}u_{241}}\ g_{u_{242}}(x_{242})\ P_{u_{24}v_{24}u_{242}}\Pr(y_{241}\,,s_0,t_0,u_{24},v_{24})$$

$$=\left(\sum_{u_{241}=1}^{k} g_{u_{241}}(x_{241})P_{u_{24}v_{24}u_{241}}\right)\left(\sum_{u_{242}=1}^{k} g_{u_{242}}(x_{242})\ P_{u_{24}v_{24}u_{242}}\right)\Pr(y_{241}\,,s_0,t_0,u_{24},v_{24})$$

$$=\left(\sum_{u_1=1}^{k} g_{u_1}(x_{241})P_{u_{24}v_{24}u_1}\right)\left(\sum_{u_1=1}^{k} g_{u_1}(x_{242})\ P_{u_{24}v_{24}u_1}\right)\Pr(y_{241}\,,s_0,t_0,u_{24},v_{24})$$

$$=\left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(y_{241},s_0,t_0,u_{24},v_{24})$$

$$=\left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(u_{24}|s_0,t_0)\Pr(s_0,t_0,y_{241},v_{24})$$

$$= \left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(u_{24}|s_0,t_0)\Pr(s_0)\Pr(t_0)\Pr(v_{24})\Pr(y_{241})$$

$$= \left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(u_{24}|s_0,t_0)\Psi_{s_0}\Psi_{t_0}\Psi_{v_{24}}\Pr(y_{241})$$

$$= \left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(u_{24}|s_0,t_0)\Psi_{s_0}\Psi_{t_0}\Psi_{v_{24}}\sum_{v_{241}=1}^{k}\Pr(y_{241},v_{241})$$

$$= \left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(u_{24}|s_0,t_0)\Psi_{s_0}\Psi_{t_0}\Psi_{v_{24}}\sum_{v_{241}=1}^{k}\Pr(y_{241}|v_{241})\Pr(v_{241})$$

$$= \left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(u_{24}|s_0,t_0)\Psi_{s_0}\Psi_{t_0}\Psi_{v_{24}}\sum_{v_{241}=1}^{k} g_{v_{241}}(y_{241})\,\Psi_{v_{241}} \tag{2.7}$$

Sustituyendo (2.7) en (2.6) y (2.6) en (2.5), se tiene que

$$\Pr(x_2,y_2,x_{21},x_{22},x_{23},x_{24},y_{24},x_{241},y_{241},x_{242}) =$$

$$= \sum_{s_0=1}^{k} g_{s_0}(x_2)\sum_{t_0=1}^{k} g_{t_0}(y_2)\left[\prod_{i=1}^{3}\left(\sum_{u=1}^{k} g_u(x_{2i})P_{s_0t_0u}\right)\right]\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\sum_{v_{24}=1}^{k} g_{v_{24}}(y_{24})$$
$$\left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} g_{u_1}(x_{24j})P_{u_{24}v_{24}u_1}\right)\right]\Pr(u_{24}|s_0,t_0)\boldsymbol{\Psi}_{s_0}\boldsymbol{\Psi}_{t_0}\boldsymbol{\Psi}_{v_{24}}\sum_{v_{241}=1}^{k} g_{v_{241}}(y_{241})\,\boldsymbol{\Psi}_{v_{241}}. \tag{2.8}$$

Reacomodando, se tiene que (2.8) es igual a

$$\sum_{s_0=1}^{k}\Psi_{s_0}\ g_{s_0}(x_2)\sum_{t_0=1}^{k}\Psi_{t_0}\ g_{t_0}(y_2)\left[\prod_{i=1}^{3}\left(\sum_{u=1}^{k} P_{s_0t_0u}\ g_u(x_{2i})\right)\right]\sum_{u_{24}=1}^{k}\Pr(x_{24}|u_{24})\ \Pr(u_{24}|s_0,t_0)$$
$$\sum_{v_{24}=1}^{k}\Psi_{v_{24}}\ g_{v_{24}}(y_{24})\left[\prod_{j=1}^{2}\left(\sum_{u_1=1}^{k} P_{u_{24}v_{24}u_1}\ g_{u_1}(x_{24j})\right)\right]\sum_{v_{241}=1}^{k}\Psi_{v_{241}}\ g_{v_{241}}(y_{241}).$$

Dado que en cada generación la probabilidad de presentar el genotipo u depende solamente de los genotipos de sus padres y en las parejas únicamente de la porción de la población con el genotipo buscado, como se mostró en (2.2) y (2.4), se pueden agrupar las probabilidades de los hermanos y sus parejas en productos, es decir,

$$\sum_{s_0=1}^{k} \Psi_{s_0}\ g_{s_0}(x_2) \sum_{t_0=1}^{k} \Psi_{t_0}\ g_{t_0}(y_2) \left[\prod_{i=1}^{3} \left(\sum_{u=1}^{k} P_{s_0 t_0 u}\ g_u(x_{2i})\right)\right] \left(\sum_{u=1}^{k} \Pr(x_{24}|u)\ \Pr(u|s_0,t_0)\right)$$
$$\sum_{v_{24}=1}^{k} \Psi_{v_{24}}\ g_{v_{24}}(y_{24}) \left[\prod_{j=1}^{2} \left(\sum_{u_1=1}^{k} P_{u_{24} v_{24} u_1}\ g_{u_1}(x_{24j})\right)\right] \sum_{v_{241}=1}^{k} \Psi_{v_{241}}\ g_{v_{241}}(y_{241}).$$

Los genotipos de los padres fundadores se denotaran por s_0 y t_0; los de la primera generación u para quien nace como miembro del pedigrí y v para quien se une al pedigrí; y, finalmente, para la segunda generación u_1 y v_1.

Con lo que finalmente se obtiene:

$$= \underbrace{\sum_{s_0=1}^{k} \Psi_{s_0}\ g_{s_0}(x_2) \sum_{t_0=1}^{k} \Psi_{t_0}\ g_{t_0}(y_2)}_{padres\ originales} \underbrace{\left[\prod_{i=1}^{4} \left(\sum_{u=1}^{k} P_{s_0 t_0 u}\ g_u(x_{2i})\right)\right] \sum_{v=1}^{k} \Psi_{v_2}\ g_{v_2}(y_{24})}_{primera\ generación}$$
$$\underbrace{\left[\prod_{j=1}^{2} \left(\sum_{u_1=1}^{k} P_{u\ v\ u_1}\ g_{u_1}(x_{24j})\right)\right] \sum_{v_1=1}^{k} \Psi_{v_1}\ g_{v_1}(y_{241}).}_{segunda\ generación}$$

2.3 Matrices de transmisión y transición genéticas

En el modelo de un locus con dos alelos, A y a, se sabe que un alelo será heredado del progenitor que pertenece al pedigrí y el otro alelo será otorgado por la pareja de dicho progenitor. Por lo tanto, para cada hijo de la pareja se tienen las siguientes cuatro opciones que distinguen que alelo da cada progenitor: AA, Aa, aA, aa.

Para conocer la probabilidad de obtener un genotipo distinguiendo con que alelo contribuirá cada progenitor, denotamos lo siguiente:

La delta de Kronecker es definida por:

$$\delta_{uv} \begin{cases} 1 & \text{si } u{=}v \\ 0 & \text{si } u{\neq}v. \end{cases}$$

Por lo que la probabilidad de que un padre con genotipo *ij*, donde *i* y *j* son los alelos que forman el genotipo, transmita el alelo *k* a su descendencia se puede escribir como:

$$\tau(k\,|i,j) = \Pr(\textit{descendiente hereda alelo } k\,|\,\textit{progenitor tiene genotipo } ij) = \tau_{ij\,k}$$

$$\tau_{ij\,k} = \frac{1}{2}(\delta_{ik} + \delta_{jk}).$$

De esta forma, se puede conocer la probabilidad de que un individuo presente cierto genotipo dados los genotipos de sus padres.

$$\Pr(uu'|ss',tt') = \Pr(\textit{tener genotipo } uu' \textit{ dados los genotipos de progenitores } ss' \textit{ y } tt')$$

$$= P_{ss'tt'uu'}$$

$$= \tau(u|ss')\ \tau(\mathrm{u}'|\mathrm{tt}') \quad = \quad \tau_{ss'u}\ \tau_{tt'u}.$$

Por ejemplo, si un progenitor tiene genotipo (AA) y su pareja (Aa) entonces:

$$\begin{aligned}
\Pr(Aa|AA,Aa) &= \tau(A|AA)\ \tau(a|Aa) \\
&= \tau_{AA\,A}\ \ \tau_{Aa\,a} \\
&= \frac{1}{2}(1+1)\frac{1}{2}(0+1) \\
&= 1/2. \\
\Pr(aa|AA,Aa) &= \tau_{AA\,a}\ \ \tau_{Aa\,a} \\
&= \frac{1}{2}(0+0)\frac{1}{2}(0+1) \\
&= 0. \\
\Pr(AA|AA,AA) &= \tau_{AA\,A}\ \ \tau_{AA\,A} \\
&= \tfrac{1}{2}(1+1)\tfrac{1}{2}(1+1) \\
&= 1.
\end{aligned}$$

Se define el vector (p_1, p_2, p_3, p_4) como el vector correspondiente a las probabilidades de heredar los genotipos 1=AA, 2=Aa, 3=aA y 4=aa.

Entonces, se tienen las entradas de la siguiente matriz correspondientes a todas las posibles probabilidades de heredar los dos alelos donde importa que progenitor dona cada alelo.

s \ t	$1 = AA$	$2 = Aa$	$3 = aA$	$4 = aa$
$1 = AA$	(1,0,0,0)	$(\frac{1}{2}, \mathbf{\frac{1}{2}}, 0, 0)$	$(\frac{1}{2}, \frac{1}{2}, 0, 0)$	(0,1,0,0)
$2 = Aa$	$(\frac{1}{2}, 0, \frac{1}{2}, 0)$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$(0, \frac{1}{2}, 0, \frac{1}{2})$
$3 = aA$	$(\frac{1}{2}, 0, \frac{1}{2}, 0)$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$(0, \frac{1}{2}, 0, \frac{1}{2})$
$4 = aa$	(0,0,1,0)	$(0, 0, \frac{1}{2}, \frac{1}{2})$	$(0, 0, \frac{1}{2}, \frac{1}{2})$	(0,0,0,1)

De esta forma, para verificar $\Pr(Aa|AA, Aa)$ se observa la segunda coordenada del vector ubicado en la primera fila y segunda columna, con lo que se tiene que $\Pr(Aa|AA, Aa) = \Pr(2 \mid 1, 2) = 1/2$.

A su vez, la probabilidad de obtener un genotipo AA, Aa ó aa, dados los genotipos de los padres (s, t) de forma que no tiene importancia con que alelo contribuye cada padre, definida como la probabilidad de transmisión, está dada por las siguientes matrices:

$$Pr(AA|s,t) \quad P_{s\ t\ AA}$$

t\s	AA	Aa	aa
AA	1	1/2	0
Aa	1/2	1/4	0
aa	0	0	0

$$Pr(Aa|s,t) \quad P_{s\ t\ Aa}$$

	AA	Aa	aa
AA	0	1/2	1
Aa	1/2	1/2	1/2
aa	1	1/2	0

$$Pr(aa|s,t) \quad P_{s\ t\ aa}$$

	AA	Aa	aa
AA	0	1/2	1
Aa	1/2	1/2	1/2
aa	1	1/2	0

Dado que cada padre contribuye con un alelo y no importa quien dio cada uno, las entradas de las matrices se obtienen encontrando la proporción del genotipo de interés dado que se conocen los genotipos de los progenitores. Así, para el caso u=AA se tiene:

Dado	Posibles combinaciones	Probabilidad de transición
s = AA, t = AA	$(\boldsymbol{A}, \boldsymbol{A}), (\boldsymbol{A}, \boldsymbol{A}), (\boldsymbol{A}, \boldsymbol{A})$ $\boldsymbol{y}$ $(\boldsymbol{A}, \boldsymbol{A})$	$Pr(AA\|AA, AA) = \mathbf{4/4}$
s = AA, t = AA	$(\boldsymbol{A}, \boldsymbol{A}), (\boldsymbol{A}, \boldsymbol{A}), (a, A)$ y (a, A)	$Pr(AA\|Aa, AA) = 1/2$
s = AA, t = AA	$(a, A), (a, A), (a, A)$ y (a, A)	$Pr(AA\|aa, AA) = 0$

$\mathrm{s} = \mathrm{AA}, \mathrm{t} = \mathrm{AA}$	$(\boldsymbol{A}, \boldsymbol{A}), (A, a), (\boldsymbol{A}, \boldsymbol{A})\ y\ (A, a)$	$Pr(AA\|AA, Aa) = 1/2$
$\mathrm{s} = \mathrm{AA}, \mathrm{t} = \mathrm{AA}$	$(\boldsymbol{A}, \boldsymbol{A}), (A, a), (a, A)\ y\ (a, a)$	$Pr(AA\|Aa, Aa) = 1/4$
$\mathrm{s} = \mathrm{AA}, \mathrm{t} = \mathrm{AA}$	$(a, A), (a, a), (a, A)\ y\ (A, a)$	$Pr(AA\|aa, Aa) = 0$
$\mathrm{s} = \mathrm{AA}, \mathrm{t} = \mathrm{AA}$	$(A, a), (A, a), (A, a)\ y\ (A, a)$	$Pr(AA\|AA, aa) = 0$
$\mathrm{s} = \mathrm{AA}, \mathrm{t} = \mathrm{AA}$	$(A, a), (A, a), (a, a)\ y\ (a, a)$	$Pr(AA\|Aa, aa) = 0$
$\mathrm{s} = \mathrm{AA}, \mathrm{t} = \mathrm{AA}$	$(a, a), (a, a), (a, a)\ y\ (a, a)$	$Pr(AA\|aa, aa) = 0$

3 Modelos de regresión logística

3.1 Definición y resultados preliminares

Los modelos estadísticos de regresión son útiles para describir la relación existente entre una variable respuesta y una o más variables explicativas. Existen varios tipos de modelos de regresión que pueden ser usados para describir esta relación. Entre estos, tenemos los modelos de regresión logística que son definidos a seguir de acuerdo con Montgomery et al. (2006) y Agresti (2007).

Definición 3.1 El **modelo de regresión logística** es un modelo lineal generalizado que busca explicar una variable **binaria**, llamada dependiente o respuesta, en función del comportamiento de otras, denominadas independientes o explicativas.

Observación 3.1 Los modelos de regresión logística son de suma importancia en diversas áreas del conocimiento, desde análisis epidemiológicos en la determinación de los factores de riesgo que ocasionan enfermedades hasta estudios de mercado que buscan conocer que variables influyen en la compra o no de ciertos tipos de productos, e incluso, estudios financieros para explicar los factores inherentes al otorgamiento o no de un crédito.

Ejemplo 3.1 La variable dependiente podría ser el desarrollar una enfermedad pulmonar, con la clasificación 0 si el individuo no la presenta y 1 si la presenta y las variables independientes que pueden ser los factores de riesgo, como antecedente hereditario, edad, peso, si se es fumador o no, entre otras.

Definición 3.2 La **función logística** que da lugar al modelo se define de la siguiente forma:

$$f(x) = \frac{1}{1 + e^{-x}} \ , \qquad x \in \mathbb{R},$$

y tiene la forma gráfica siguiente (ver código en R en el apéndice de esta tesina):

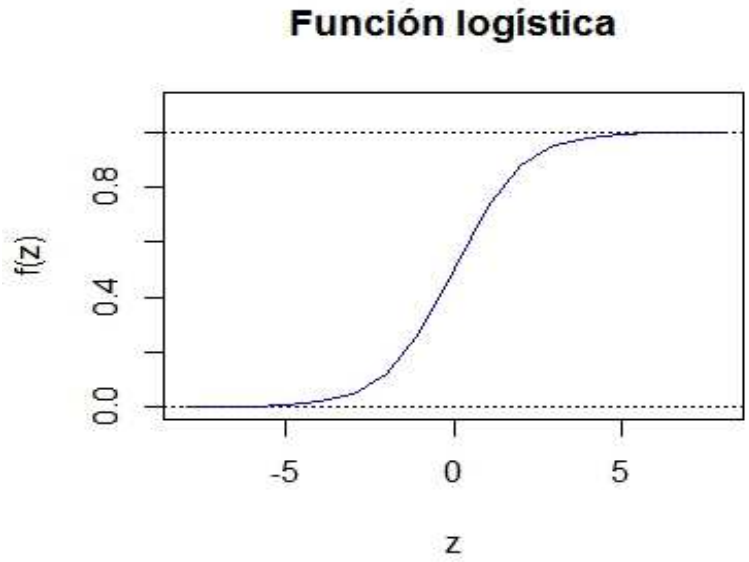

Figura 2. Gráfica de la función logística $f(x) = \frac{1}{1+e^{-x}}, x \in \mathbb{R}$. (Elaboración propia)

Analizando los valores que toma la función, se tiene que el dominio son todos los reales y el rango es el intervalo cerrado de 0 a 1. De esta forma, los valores estimados obtenidos en el modelo de regresión logística siempre se encuentran entre 0 y 1. Esto no necesariamente es cierto en otros modelos de regresión.

La forma de la función indica que el efecto combinado de las variables explicativas será bajo si sus valores lo son, y aumentará conforme ellas crezcan, siempre manteniendo el límite dado por el rango de la función.

3.2 Modelos de regresión logística

Casella et. al (2002), Montgomery et al. (2006) y Agresti (2007) exponen la conformación del modelo. A partir de dicha bibliografía se describe a continuación la generalización para cualquier cantidad de variables explicativas.

Se consideran las variables

$X_1,\ X_2, \ldots, X_n$: llamadas las variables explicativas y

Y: la variable respuesta a ser obtenida a partir de las explicativas.

Suponga que se puede escribir Y de la siguiente forma:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \alpha + \sum_{i=1}^{n} \beta_i X_i$$

donde $\alpha \in \mathbb{R}$ y $\beta_i \in \mathbb{R}$, $i = 1, \ldots, n$, son parámetros desconocidos que necesitan ser estimados mediante la información conocida de las variables explicativas a partir de la muestra o población estudiada.

Sea $X = (X_1, X_2, \ldots, X_n)$. Suponga que Y asume valores en {0,1} y que la probabilidad de obtener un éxito en Y (por ejemplo, la presencia de una enfermedad) sea dada por:

$$\pi(X) = P(Y = 1 | X_1, X_2, \ldots, X_n).$$

Por definición, Y es una variable aleatoria con distribución *Bernoulli* con parámetro $\pi(X)$. Por lo tanto, su valor esperado es la probabilidad de obtener un éxito, es decir,

$$\begin{aligned} E(Y) &= (0)\ P(Y = 0 | X_1, X_2, \ldots, X_n) + (1) P(Y = 1 | X_1, X_2, \ldots, X_n) \\ &= 0\big(1 - \pi(X)\big) + 1(\pi(X)) \\ &= 0 + \pi(X) \\ &= \pi(X). \end{aligned}$$

Asumiendo que el comportamiento de la variable respuesta respecto de las explicativas es tal que su densidad satisface a la función logística, se tiene que el **modelo logístico** se puede escribir de la siguiente forma:

$$E(Y) = \pi(X) = \frac{1}{1 + e^{-Y}}. \tag{3.1}$$

Por la hipótesis sobre la forma de Y se tiene que

$$E(Y) = \pi(X) = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{n} \beta_i X_i)}}. \tag{3.2}$$

Definición 3.3 La transformación **logit** está dada por:

$$logit[p] = \log\left(\frac{p}{1-p}\right),$$

donde $p \in (0,1)$ es la probabilidad de ocurrencia del fenómeno estudiado, es decir, para Y que asume los valores 0 y 1 entonces $p = \Pr(Y = 1)$.

Entonces, mediante la transformación logit, se linealiza (3.2) de la siguiente forma:

$$\begin{aligned}
logit[\pi(X)] = log\left(\frac{\pi(X)}{1-\pi(X)}\right) &= log\left(\frac{\dfrac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}{1-\dfrac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}\right) \\
&= log\left(\frac{\dfrac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}{\dfrac{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}-1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}\right) \\
&= log\left(\frac{\dfrac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}{\dfrac{e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}\right) \\
&= log\left(\frac{1}{e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}\right) \\
&= log\left(e^{(\alpha+\sum_{i=1}^{n}\beta_i X_i)}\right) \\
&= \alpha + \sum_{i=1}^{n}\beta_i X_i.
\end{aligned}$$

Observación 3.2 Cada parámetro β_i, $i = 1, \ldots, n$, representa el efecto que tiene su respectiva variable independiente X_i en la variable dependiente. La combinación lineal de las variables explicativas es el efecto que causan de forma conjunta $X_1,\ X_2, \ldots, X_n$ a la variable respuesta Y.

Otra interpretación que se puede dar al modelo es mediante la **razón de momios** (odds ratios), también llamada razón de posibilidades o ventaja.

Los momios de un evento son el cociente del número de veces que se espera ocurra el evento entre el número de veces que se espera no ocurra.

Ejemplo 3.2 Si los momios de desarrollar una enfermedad son 5 a 1, entonces se dice que es cinco veces más probable desarrollar esa enfermedad que no desarrollarla.

Sea p la probabilidad de que ocurra un evento determinado e indique por M a los momios, entonces tenemos la siguiente relación:

$$M = \frac{p}{1-p} \quad \text{o equivalentemente} \quad p = \frac{M}{1+M}.$$

Ejemplo 3.3 Suponiendo que se tiene la probabilidad 0.6 de manifestar cierto tipo de rasgo hereditario y 0.4 de no hacerlo, entonces los momios son $0.6/0.4 = 1.5$. Esto significa que presentar el rasgo hereditario es 1.5 veces más probable a no presentarlo. Con la relación dada, conociendo los momios podemos saber la probabilidad de éxito, si se tienen momios de 3 (3 a 1), entonces la probabilidad del evento es 0.75.

La **razón de momios o posibilidades** (odds ratio) está directamente relacionada con los parámetros de un modelo de regresión logística.

Considere lo siguiente. Sean

$$Y\colon \textit{desarrollar rasgo hereditario}$$

$$X\colon \textit{alguno de los dos progenitores tiene el rasgo}$$

variables que asumen los valores cero y uno con probabilidades de ocurrencia tales que

$$\Pr(Y = 1|X = 1) = p_1$$

es la probabilidad de desarrollar el rasgo dado que al menos un progenitor lo presenta; y

$$\Pr(Y = 1|X = 0) = p_2$$

es la probabilidad de desarrollar rasgo hereditario dado que ningún progenitor lo presenta.

Por lo tanto, los momios de desarrollar el rasgo hereditario dado que al menos uno de los progenitores tenga y dado que ninguno de ellos lo presente son, respectivamente,

$$M_1 = \frac{p_1}{1-p_1} \quad y \quad M_2 = \frac{p_2}{1-p_2},$$

y de esta forma, la **razón de posibilidades (OR)** es

$$OR = \frac{M_1}{M_2}.$$

Observación 3.3 OR siempre es positivo. Un valor igual a uno indica que las variables no se relacionan. Valores cercanos a cero o mayores a uno indican algún grado de relación entre ellas.

Si la razón de momios es igual a 5, entonces esto indica que es mucho más probable desarrollar el rasgo si este está presente en alguno de los progenitores respecto de no estarlo. Lo cual muestra una asociación fuerte entre el rasgo y que algún progenitor lo tenga.

Al usar la razón de momios en la regresión logística se puede verificar el efecto que tienen las variables explicativas en la variable repuesta.

Lema 3.1 Sea un modelo de regresión logística con solamente una variable explicativa.

a) El **momio** de que suceda el evento de interés respecto de que no ocurra es

$$M = \frac{\pi(X)}{1-\pi(X)} = e^{\alpha}e^{\beta X}. \tag{3.3}$$

b) El momio al aumentar el valor de X en una unidad es:

$$M' = \frac{\pi(X+1)}{1-\pi(X+1)} = e^{(\alpha+\beta(X+1))} = e^{\alpha}e^{\beta(X+1)} = e^{\alpha}e^{\beta X}e^{\beta}. \tag{3.4}$$

Demostración:

(a) Para demostrar (3.3) se considera la igualdad dada por (3.2), entonces

$$M = \frac{\pi(X)}{1-\pi(X)} = \left(\frac{\frac{1}{1+e^{-(\alpha+\beta X)}}}{1-\frac{1}{1+e^{-(\alpha+\beta X)}}} \right)$$

$$= \left(\frac{\frac{1}{1+e^{-(\alpha+\beta X)}}}{\frac{1+e^{-(\alpha+\beta X)}-1}{1+e^{-(\alpha+\beta X)}}} \right)$$

$$= \left(\frac{1}{e^{-(\alpha+\beta X)}} \right) = e^{(\alpha+\beta X)} = e^{\alpha} e^{\beta X}.$$

(b) De forma análoga, para (3.4)

$$M' = \frac{\pi(X+1)}{1-\pi(X+1)} = \left(\frac{\frac{1}{1+e^{-(\alpha+\beta(X+1))}}}{1-\frac{1}{1+e^{-(\alpha+\beta(X+1))}}} \right)$$

$$= \left(\frac{\frac{1}{1+e^{-(\alpha+\beta(X+1))}}}{\frac{1+e^{-(\alpha+\beta(X+1))}-1}{1+e^{-(\alpha+\beta(X+1))}}} \right)$$

$$= \left(\frac{1}{e^{-(\alpha+\beta X)}} \right) = e^{(\alpha+\beta(X+1))}$$

$$= e^{\alpha} e^{\beta(X+1)} = e^{\alpha} e^{\beta X} e^{\beta}.$$

Observación 3.4 Del lema 3.1 se puede observar que al aumentar una unidad en la variable explicativa, aumentan de manera multiplicativa en e^{β} las posibilidades de ocurrencia del evento de interés con respecto de que no ocurra.

Entonces, la razón de momios es

$$\frac{M'}{M} = \frac{e^{\alpha} e^{\beta X} e^{\beta}}{e^{\alpha} e^{\beta X}} = e^{\beta},$$

Por lo tanto, se puede ver que si $\beta = 0$, entonces $e^{\beta} = 1$. Esto quiere decir que la variable explicativa no es significativa al modelo. Si $\beta \neq 0$, la variable explicativa tiene algún grado de relación con la variable respuesta.

Observación 3.5 Si se toma en cuenta la expresión logit del modelo, se tiene una relación lineal en la cual $logit[\pi(X)]$ incrementa en β unidades por cada cambio de una unidad en X.

Lema 3.2 Suponga que existen dos o más variables explicativas $X_1, X_2, \dots, X_n$. De esta forma, el modelo general en su forma logit es:

$$\text{(a)} \qquad log\Big(\tfrac{\pi(X)}{1-\pi(X)}\Big) = \alpha + \textstyle\sum_{i=1}^{n} \beta_i X_i \tag{3.5}$$

$$(b) \qquad e^{log\left(\frac{\pi(X)}{1-\pi(X)}\right)} = e^{\alpha} \prod_{i=1}^{n} e^{(\beta_i)Xi} \tag{3.5}$$

Demostración:

(a) Para demostrar (3.5) se usa (3.2) de la siguiente manera

$$\begin{aligned} log\left(\frac{\pi(X)}{1-\pi(X)}\right) &= log\left(\frac{\dfrac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}{1-\dfrac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}\right) \\ &= log\left(\frac{\dfrac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}{\dfrac{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}-1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}}\right) \\ &= log\left(\frac{1}{e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}\right) \\ &= log\big(e^{(\alpha+\sum_{i=1}^{n}\beta_i X_i)}\big) = \alpha + \sum_{i=1}^{n}\beta_i X_i. \end{aligned}$$

(b) Para demostrar (3.6), a partir de (3.5) se tiene

$$e^{log\left(\frac{\pi(X)}{1-\pi(X)}\right)} = e^{\alpha+\sum_{i=1}^{n}\beta_i X_i} = e^{\alpha}\prod_{i=1}^{n} e^{(\beta_i)Xi} \quad .$$

De (3.5) se interpreta que los parámetros β_i representan el efecto de su respectiva variable explicativa X_i en el logaritmo de los momios.

En (3.6), se tiene que e^{β_i} es el efecto multiplicativo en las posibilidades por cada unidad de cambio en X_i cuando se mantienen las demás variables explicativas controladas o se suponen constantes.

3.3 Estimación de parámetros de un modelo de regresión logística

Existen varias formas de estimar los parámetros de una regresión logística. Algunas de ellas serán vistas a seguir.

3.3.1 Estimadores de máxima verosimilitud

De acuerdo con Montgomery et al. (2006) la mejor forma para obtener los estimadores es mediante la estimación por máxima verosimilitud. Dicha estimación maximiza la probabilidad de obtener el conjunto de datos observados.

Definición 3.4 Sean $Y_1, Y_2, \dots, Y_n$ observaciones independientes cuyos resultados pueden ser ceros o unos con una distribución Bernoulli(p), $p \in (0,1)$ y sean $X_1, X_2, \dots, X_n$ las variables explicativas. La forma general del modelo de regresión logística para cada observación Y_i es:

$$E(Y_i) = \frac{e^{(\alpha+\sum_{i=1}^{n}\beta_i X_i)}}{1+e^{(\alpha+\sum_{i=1}^{n}\beta_i X_i)}} = \pi_i, \qquad i = 1,2,\dots,n.$$

Por hipótesis, cada observación Y_i tiene distribución Bernoulli(p), con $p = \pi_i$ para $i = 1,2,\dots,n$. Por lo tanto

$$f_i(Y_i) = {\pi_i}^{Y_i}(1-\pi_i)^{1-Y_i}\ , \qquad i = 1,2,\dots,n.$$

De esta forma, por la sección 1.2, la función de verosimilitud es el producto de las n funciones de densidad, es decir, si $\tau = (\alpha, \beta_1, \beta_2, \dots, \beta_n)$, entonces

$$\begin{aligned} L(Y_1, Y_2, \dots, Y_n|\tau) &= L(Y_1, Y_2, \dots, Y_n|\alpha, \beta_1, \beta_2, \dots, \beta_n) \\ &= \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} {\pi_i}^{Y_i}(1-\pi_i)^{1-Y_i}. \end{aligned}$$

Dado que se busca maximizar la función y es más flexible trabajar el logaritmo de la función de verosimilitud, es decir, la log-verosimilitud, la cual conserva la igualdad de los valores tomados en los puntos máximos, se tiene

$$\begin{aligned} log\ L(Y_1, Y_2, \dots, Y_n; \tau) &= log\left(\prod_{i=1}^{n} f_i(Y_i)\right) \\ &= log\left(\prod_{i=1}^{n} {\pi_i}^{Y_i}(1-\pi_i)^{1-Y_i}\right) \\ &= \sum_{i=1}^{n} log({\pi_i}^{Y_i}(1-\pi_i)^{1-Y_i}) \\ &= \sum_{i=1}^{n} log\left[\frac{{\pi_i}^{Y_i}(1-\pi_i)}{(1-\pi_i)^{Y_i}}\right] \\ &= \sum_{i=1}^{n} Y_i\ log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \\ &= \sum_{i=1}^{n}\left[Y_i\ log\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n}\log(1-\pi_i)\,. \end{aligned}$$

Por hipótesis se tiene que

$$1-\pi_i = \frac{1}{\left[1+e^{(\alpha+\sum_{i=1}^{n}\beta_i X_i)}\right]}.$$

Por lo tanto,

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{i=1}^{n}\beta_i X_i.$$

De ésta forma, la log-verosimilitud se puede expresar como

$$\log L(Y_1, Y_2, \dots, Y_n|\tau) =$$

$$= \sum_{i=1}^{n} Y_i\left(\alpha + \sum_{i=1}^{n}\beta_i X_i\right) - \sum_{i=1}^{n}\log\left[1+e^{(\alpha+\sum_{i=1}^{n}\beta_i X_i)}\right].$$

Los estimadores serán las soluciones de las derivadas parciales igualadas a 0 respecto de cada parámetro.

El cálculo de los estimadores de máxima verosimilitud se realiza mediante programas de cómputo usando métodos numéricos o el algoritmo de mínimos cuadrados iterativamente reponderados (IRLS) tratado a detalle por Montgomery et al. (2006).

Conociendo los estimadores resultantes $\hat{\alpha}, \widehat{\beta_1}, \widehat{\beta_2}, \dots, \widehat{\beta_n}$, el **valor esperado del modelo de regresión logística** se escribe:

$$E(Y_i) = \pi_i = \frac{1}{1 + e^{-\left(\hat{\alpha} + \sum_{i=1}^{n} \widehat{\beta_i} X_i\right)}}.$$

3.3.2 Intervalos de confianza

Agresti (2007), señala que un intervalo de confianza para el parámetro β en el modelo de regresión logística, es:

$$\hat{\beta} \pm z_{\alpha/2}(SE),$$

donde $\hat{\beta}$ es el valor estimado del parámetro,

$z_{\alpha/2}$ se obtiene de las tablas de una normal estándar y representa el nivel de confianza (si tomamos 95% de nivel de confianza se tiene $z_{0.975} = 1.96$) y

SE es el error estándar estimado.

Al tener el intervalo de confianza del parámetro, es posible obtener también el intervalo de confianza del efecto del mismo usando los límites obtenidos $(e^{\hat{\beta} - z_{\alpha/2}(SE)}, e^{\hat{\beta} + z_{\alpha/2}(SE)})$, con lo cual se puede inferir con un 95% de confianza el efecto de la variable explicativa en el modelo.

3.3.3 Prueba de significancia

Como se trató anteriormente, mediante la razón de momios se puede establecer la significancia de las variables lo cual se complementa con el estadístico de prueba, descrito en el capítulo quinto de Agresti (2007), que tiene la hipótesis nula que señala que la probabilidad de que ocurra el evento de interés es independiente a X si H_0: $\beta = 0$.

El estadístico

$$z = \frac{\hat{\beta}}{SE}$$

tiene una distribución normal estándar cuando $\beta = 0$ por lo que se procede a observar el valor del estadístico y el comportamiento del p-valor para rechazar o no la hipótesis.

3.3.4 Desviación

Esta es una prueba que muestra si el modelo propuesto es adecuado o no. Toma en cuenta la log-verosimilitud del modelo ajustado con p parámetros y la de un modelo que representa perfectamente a la muestra de tamaño n y que tiene los n parámetros. Para $p < n$ el valor de la log-verosimilitud del modelo completo es mayor a la del ajustado dado que éste último cuenta con menos parámetros.

Para fines prácticos, se tomará en cuenta a la log-verosimilitud del modelo completo como $\log L(\beta)$ y a la log-verosimilitud del modelo ajustado como $logL(\hat{\beta})$, con lo cual Montgomery et al. (2006), en el capítulo 13, define la **desviación** como:

$$desv(\alpha, \beta_1, \beta_2, \dots, \beta_n) = 2 \ln L(\beta) - 2\ lnL\left(\hat{\beta}\right).$$

Si el modelo es correcto y la muestra es grande (suficiente), entonces la desviación del modelo tiene una distribución Ji-cuadrada con *(n-p)* grados de libertad, con los siguientes mecanismos para decidir:

Si:

$desv(\alpha, \beta_1, \beta_2, \dots, \beta_n) \leq X^2_{\alpha, n-p}$ entonces el modelo propuesto es adecuado.

$desv(\alpha, \beta_1, \beta_2, \dots, \beta_n) > X^2_{\alpha, n-p}$ entonces el modelo propuesto no es adecuado.

De forma análoga, se pueden realizar pruebas de hipótesis sobre subconjuntos de parámetros tomando el modelo completo y contrastándolo con uno reducido.

4 Modelos de regresión logística aplicados a genética

4.1 Consideraciones preliminares

El estudio de la presencia de enfermedades o rasgos, que pueden ser clasificados como datos binarios, en los que se sospecha que existe alguna relación entre la herencia biológica y la observación o no de dichas características ha sido posible gracias al uso de diversas metodologías, entre las que destacan las estadísticas.

El modelo de regresión logística es una de dichas metodologías estadísticas. Este tipo de modelo permite conocer los factores de riesgo que tienen algún grado de influencia en presentar o no una enfermedad o característica. Entre los factores que pueden intervenir, se destaca la herencia biológica determinada por el genotipo de los individuos. Asimismo, de acuerdo con la enfermedad estudiada, se pueden utilizar otras variables en combinación con el genotipo, para así explicar de la mejor forma posible la causalidad de la característica analizada. Algunas de estas variables explicativas de interés son la edad, el género, ser fumador o no y los niveles de glucosa del individuo.

Para llevar a cabo la implementación del modelo, Bonney (1986) propone un modelo general el cual se puede ajustar conforme a la información con que se cuente y las variables que se quieran considerar. A continuación se disertará tanto sobre el modelo general como de los específicos derivados de este, estudiados en Bonney (1986).

4.2 Modelo general

El modelo general considera que entre los n individuos de una población se tiene algún tipo de relación o dependencia, la cual tiene algún papel en el desarrollo la enfermedad o característica estudiada. Los modelos en que la dependencia es desconocida o es por alguna relación biológica, son casos específicos del modelo general y se expondrán más adelante.

Las observaciones de los individuos se denotan $Y = (Y_1, Y_2, \dots, Y_n)$ donde, para $i = 1,2, \dots, n$,

$$Y_i = \begin{cases} 0 & \text{si no presenta el rasgo de interés} \\ 1 & \text{si presenta el rasgo .} \end{cases}$$

Se asumirá que la variable o variables explicativas están relacionadas con su respectivo *i*-ésimo individuo. Así cada variable explicativa se denotará por X_i y estará asociada al rasgo Y_i del *i*-ésimo individuo.

El vector de variables explicativas será denotado por

$$X = (X_1, X_2, \dots, X_n).$$

Entonces, la probabilidad de presentar el rasgo de interés dado el vector de variables explicativas se denotará, para próximos desarrollos, por:

$$\Pr(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | X_1 = x_1, X_2 = x_2, \dots, X_n = x_x) = \Pr(Y|X),$$

con lo que se tiene que:

$$\begin{aligned} \Pr(Y|X) &= \Pr(Y_1, Y_2, \dots, Y_n|X) \\ &= \Pr(Y_1|X) \Pr(Y_2, Y_3, \dots, Y_n|Y_1, X) \\ &= \Pr(Y_1|X) \Pr(Y_2|Y_1, X) \Pr(Y_3, \dots, Y_n|Y_1, Y_2, X). \end{aligned}$$

De forma análoga, usando propiedades de probabilidad condicional, se llega a:

$$\Pr(Y|X) = \Pr(Y_1|X) \Pr(Y_2|Y_1, X) \Pr(Y_3|Y_1, Y_2, X) \dots \Pr(Y_n|Y_1, Y_2, \dots, Y_{n-1}, X). \tag{4.1}$$

Debido a que el *i*-ésimo individuo sólo está relacionado a la *i*-ésima variable explicativa, entonces cada factor de (4.1) estará dado por:

$$\Pr(Y_i|Y_1, Y_2, \dots, Y_{i-1}, X) = \Pr(Y_i|Y_1, Y_2, \dots, Y_{i-1}, X_i).$$

Para describir la presencia o no del rasgo de interés, Bonney (1986) define una variable que ha sido modificada en esta tesina y se muestra a continuación:

$$Z_i = \begin{cases} -1 & \text{si } \mathrm{Y_i} = 0 \\ 0 & \text{si } \mathrm{Y_i} \text{ desconocida} \\ 1 & \text{si } \mathrm{Y_i} = 1. \end{cases}$$

Para describir el modelo de regresión logística, se definirán las n transformaciones *logit* (tema expuesto en el capítulo 3) siguientes:

$$\omega_1 = \alpha + \beta X_1$$

$$\omega_2 = \alpha + \gamma_1 Z_1 + \beta X_1$$

$$\vdots$$

$$\omega_n = \alpha + \gamma_1 Z_1 + \gamma_2 Z_2 + \cdots + \gamma_{n-1} Z_{n-1} + \beta X_1,$$

es decir, (Bonney, 1986)

$$\begin{aligned} \omega_1 &= \alpha + \beta X_1 \\ \omega_k &= \alpha + \sum_{i=1}^{k-1} \gamma_i Z_i + \beta X_1 \ , \qquad con\ k = 2,3,\dots,n. \end{aligned} \tag{4.2}$$

donde α, β y γ_i, $i = 1,2,\dots,n-1$ son parámetros desconocidos que necesitan ser estimados y pueden tomar valores en el intervalo $(-\infty, \infty)$.

Proposición 4.1: Usando la transformación logit, se tiene que (4.1) puede ser escrita como

$$\Pr(Y|X) = \prod_{i=1}^{n} \left[\frac{e^{\omega_i Y_i}}{1 + e^{\omega_i}} \right] \tag{4.3}$$

donde el producto es sobre $\mathrm{Y_i}$ observadas.

Demostración: Por (4.1) se tiene que

$$\Pr(Y|X) = \Pr(Y_1|X)\Pr(Y_2|Y_1, X)\Pr(Y_3|Y_1, Y_2, X)\dots\Pr(Y_n|Y_1, Y_2, \dots, Y_{n-1}, X).$$

Por lo tanto, asumiendo que Y es observado y tiene un comportamiento logístico (visto en el capítulo 3), y mediante las n transformaciones logit (4.2) se tiene que la probabilidad conjunta es el producto de las probabilidades de ocurrencia individuales, es decir,

$$\Pr(Y|X) = \left(\frac{e^{\omega_1 Y_1}}{1+e^{\omega_1}}\right)\left(\frac{e^{\omega_2 Y_2}}{1+e^{\omega_2}}\right)\ldots\left(\frac{e^{\omega_n Y_n}}{1+e^{\omega_n}}\right)$$

$$= \prod_{i=1}^{n}\left[\frac{e^{\omega_i Y_i}}{1+e^{\omega_i}}\right].$$

De acuerdo con (4.3), se puede notar de forma preliminar en que magnitud intervienen en la probabilidad conjunta las características vistas de los individuos. Es decir, supongamos que el individuo Y_j precede a Y_i por lo que Y_i depende de Y_j $(j < i)$. Entonces, al presentar Y_j la enfermedad $(Z_j = 1)$ los momios o posibilidades de que Y_i presente el rasgo aumentan en e^{γ_j}. Por otro lado, si Y_j no presenta la enfermedad $(Z_j = -1)$ entonces los momios de que Y_i presente la característica disminuyen en e^{γ_j}. Si no se conoce el valor de la observación Y_j $(Z_j = 0)$ entonces los momios de Y_i no tienen ningún cambio.

Observaciones: 4.1 Cada unidad de incremento o decremento en la variable explicativa (X_i), aumenta o disminuye los momios en e^{β}.

4.2 Se puede notar que el valor que tomarán los parámetros al ser estimados es la magnitud de cambio en su respectiva variable.

4.3 La interpretación de los momios se trató en el capítulo 3.

4.3 Modelos específicos

Se retomarán los tipos de modelos propuestos por Bonney (1986) derivados de adecuaciones al modelo general.

(a) Modelos en que no hay dependencia basada en relaciones biológicas

Son útiles cuando es muy laborioso realizar el modelo general y las observaciones no tienen relación entre si o no están bien agrupadas de acuerdo a la relación familiar.

(a.1) Modelo equitativamente predictivo.

En esta versión del modelo, los parámetros $\gamma_i,\ i = 1,2,\dots,n-1$, que miden factores de cambio de todos los individuos son iguales, con lo cual, las trasformaciones logit del modelo general son:

$$\omega_i = \alpha + \gamma Z_1 + \gamma Z_2 + \cdots + \gamma Z_{i-1} + \beta X_i\ , \qquad i = 2,3,\dots,n$$

es decir,

$$\omega_i = \alpha + \gamma \sum_{j=1}^{i-1} Z_j\ + \beta X_i\ , \qquad i = 2,3,\dots,n.$$

El caso ω_1 es el mismo que en el modelo general.

Por lo tanto, la relación (4.1) se mantiene como en el caso general y las transformaciones logit muestran que existe la misma magnitud de cambio por cada individuo predecesor existente. Con ello, queda definido el modelo.

(a.2) Modelo de dependencia en serie.

Tomando en cuenta las observaciones y la estructura de dependencia familiar de orden 1, es decir, que los individuos dependen únicamente de su predecesor y no de todos los anteriores, se tiene que la transformación logit ω_1 será como en el caso general y para las demás se tendrá

$$\omega_i = \alpha + \gamma Z_{i-1} + \beta X_i\ , \qquad i = 2,3,\dots,n.$$

De esta forma, (4.1) se transforma en

$$\begin{aligned}\Pr(Y|X) &= \Pr(Y_1|X)\Pr(Y_2|Y_1,X)\Pr(Y_3|Y_2,X)\dots\Pr(Y_n|Y_{n-1},X)\\ &= \Pr(Y_1|X_1)\prod_{i=2}^{n}\Pr(Y_i|Y_{i-1},X_i)\\ &= \prod_{i=1}^{n}\left[\frac{e^{\omega_i Y_i}}{1+e^{\omega_i}}\right].\end{aligned}$$

(b) Modelos con dependencia basada en relaciones biológicas

Este tipo de modelos asume que las observaciones están ordenadas de forma que los padres preceden a su descendencia y que los hijos de cada generación se ordenan de acuerdo con su orden de nacimiento.

Para cualquier individuo *i,* las letras *P, M y C* denotarán a su respectivo padre, madre y pareja. Entre los hermanos del individuo *i*, *B(1)* denotará al primogénito y *B(-1),* al hermano que precede a *i*.

(b.1) Modelos clase A

Los modelos clase A suponen que la única dependencia entre los individuos es entre padres e hijos. Entonces, dada la información de los padres del individuo *i*, las demás observaciones no contribuyen en brindar más explicación a la presencia o no del rasgo de interés.

La probabilidad de que cada individuo presente la característica es

$$\Pr(Y_i|Y_1, Y_2, \dots, Y_{i-1}, X_i) = \Pr(Y_i|Y_P, Y_M, X_i),$$

y las transformaciones logit se definen de la forma

$$\omega_i = \alpha + \gamma_P Z_P + \gamma_M Z_M + \beta X_i, \qquad i = 1,2, \dots, n,$$

notando que si la información del padre o la madre es desconocida, por definición Z_P *o* Z_M tomarán el valor 0.

Entonces, el modelo queda definido de la siguiente manera:

$$\Pr(Y|X) = \prod_{i=1}^{n} \Pr(Y_i|Y_P, Y_M, X_i) = \prod_{i=1}^{n} \left[\frac{e^{\omega_i Y_i}}{1 + e^{\omega_i}}\right]$$

(b.2) Modelos clase B

En el caso de los modelos clase B, se asume que los progenitores y el primogénito brindan la información que determina la presencia o no de la característica de interés en el individuo estudiado. Se tiene entonces para el *i*-ésimo individuo de (4.1) que

$$\Pr(Y_i|Y_1, Y_2, \dots, Y_{i-1}, X_i) = \Pr(Y_i|Y_P, Y_M, Y_{B(1)}, X_i),$$

con lo cual las transformaciones logit son

$$\omega_i = \alpha + \gamma_P Z_P + \gamma_M Z_M + \gamma_{B(1)} Z_{B(1)} + \beta X_i,$$

con la aclaración de que si *i* es el primogénito entonces no existe uno que lo preceda y por lo tanto $Z_{B(1)} = 0$, y si la información del padre o la madre es desconocida, por definición Z_P *o* Z_M tomarán el valor 0.

Entonces, el modelo queda definido por

$$\Pr(Y|X) = \prod_{i=1}^{n} \Pr(Y_i|Y_P, Y_M, Y_{B(1)}, X_i) = \prod_{i=1}^{n} \left[\frac{e^{\omega_i Y_i}}{1+e^{\omega_i}}\right].$$

El modelo se puede extender a que los padres y los primeros *r* hermanos brindan la información determinante.

(b.3) Modelos clase C

Estos modelos consideran que los padres y el hermano predecesor del individuo *i* contribuyen en alguna magnitud con la presencia o no del rasgo de interés en el individuo *i*. Para cada individuo se tiene que la probabilidad de presentar el rasgo de interés es

$$\Pr(Y_i|Y_1, Y_2, \ldots, Y_{i-1}, X_i) = \Pr(Y_i|Y_P, Y_M, Y_{B(-1)}, X_i)$$

con lo cual las transformaciones logit son:

$$\omega_i = \alpha + \gamma_P Z_P + \gamma_M Z_M + \gamma_{B(-1)} Z_{B(-1)} + \beta X_i, \qquad i = 1,2,\ldots,n,$$

con la consideración de que si *i* es el primogénito o no se cuenta con información sobre el hermano que lo precede entonces $Z_{B(-1)} = 0$, y si la información del padre o la madre es desconocida, por definición Z_P *o* Z_M tomarán el valor 0.

Entonces, con las transformaciones dadas, el modelo tiene la forma

$$\Pr(Y|X) = \prod_{i=1}^{n} \Pr(Y_i|Y_P, Y_M, Y_{B(-1)}, X_i) = \prod_{i=1}^{n} \left[\frac{e^{\omega_i Y_i}}{1+e^{\omega_i}}\right].$$

Al igual que en los modelos clase B, se puede tener un mayor orden de dependencia.

Observación 4.4 Se puede hacer uso de combinaciones entre modelos biológicos y no biológicos dependiendo del estudio y de la información con la que se cuenta.

4.4 Modelos de regresión logística aplicados al análisis genético

Los modelos hasta ahora expuestos son buenos para el análisis de riesgo de presentar ciertos rasgos de interés, pero para el análisis genético se necesita algo más. El análisis genético busca explicar los mecanismos de segregación de los genes y con ello la manifestación de los fenotipos entre generaciones.

Supongamos como en el capítulo 2, que cada individuo tiene uno de los k genotipos posibles. También, como en el capítulo 2, sea $g = (g_1, g_2, \dots, g_n)$ el vector de los genotipos correspondientes a n individuos en un pedigrí. Si g es conocido entonces se le puede considerar en el modelo como variable explicativa. Sin embargo, la mayoría de las veces g es un vector aleatorio y la obtención de la probabilidad de su ocurrencia es de interés fundamental.

En el modelo general presentado en la sección 4.2 con una variable explicativa y considerando la trasmisión de genotipos entre generaciones se tiene:

$$\Pr(g, Y|X) = \Pr(g)\Pr(Y|g, X).$$

Al ser el vector g desconocido, por la ley de probabilidad total, se puede escibir

$$\Pr(Y|X) = \sum_{g=1}^{k} \Pr(g)\Pr(Y|g, X). \tag{4.4}$$

En el caso en que el estudio no tenga variables explicativas, el modelo se reduce a:

$$\Pr(g, Y) = \Pr(g)\Pr(Y|g)$$

como fue visto en el capítulo 2.

Al ser el vector g desconocido, por la ley de probabilidad total, se tiene:

$$\Pr(Y) = \sum_{g=1}^{k} \Pr(g)\Pr(Y|g).$$

Para conocer la distribución conjunta de los genotipos, se retomará lo visto en el capítulo 2. De esta forma, P_{stu} es la probabilidad de que el hijo tenga genotipo u dado que los genotipos de los padres son s y t, es decir,

$$P_{stu} = \Pr(u \mid s, t).$$

Además, Ψ_v es la probabilidad de que un individuo en la población tenga genotipo v.

Entonces, de lo desarrollado en el capítulo 2, se tiene que la probabilidad conjunta de observar un genotipo en n individuos es:

$$\begin{aligned}\Pr(g) = \Pr(g_1, g_2, \dots, g_n) &= \Pr(g_1)\Pr(g_2|g_1) \dots \Pr(g_n|g_1, g_2, \dots, g_{n-1}) \\ &= \prod_{i=1}^{n} p_i\end{aligned}$$

donde

$$\begin{aligned}p_i &= \Pr(g_i|g_1, g_2, \dots, g_{i-1}) \\ &= \begin{cases} \Psi_{g_i} & \textit{si los progenitores son desconocidos} \\ P_{g_P\, g_M\, g_i} & \textit{si los progenitores son conocidos} \end{cases}.\end{aligned}$$

Para calcular dichas probabilidades, se hará uso de las matrices de transición genéticas vistas en el capítulo 2.

Dados los genotipos mediante las matrices de transición y las variables explicativas, la probabilidad condicional de presentar el fenotipo, rasgo o enfermedad es

$$\Pr(Y|g, X) = \Pr(Y_1|g, X) \prod_{i=2}^{n} \Pr(Y_i|g, Y_1, Y_2, \dots, Y_{i-1}, X).$$

Del capítulo 2 sabemos que dado el genotipo de un individuo, su fenotipo es independiente de los demás y este genotipo solamente determina la posibilidad de presentar el fenotipo de ese mismo individuo. Por lo tanto, se puede escribir

$$\Pr(Y|g, X) = \Pr(Y_1|g_1, X) \prod_{i=2}^{n} \Pr(Y_i|g_i, Y_1, Y_2, \dots, Y_{i-1}, X). \tag{4.5}$$

Con los debidos ajustes al modelo general descritos en la sección 4.2, simplemente se añade una nueva variable explicativa g_i para cada Y_i.

Por ejemplo, en los modelos clase A, la transformación logit se puede definir de la siguiente forma:

$$\omega_i(g_i) = \alpha_{g_i} + \gamma_P Z_P + \gamma_M Z_M + \beta X_i$$

con $\omega_1(g_1) = \alpha_{g_1} + \beta X_1$.

Solamente α depende de g_i. Esto significa que el genotipo no interactúa con los demás factores tomados en cuenta. Cabe mencionar que los demás parámetros pueden depender de g_i si el investigador lo cree necesario. El ajuste en cada tipo de modelo es homólogo.

Ahora bien, usando (4.4) y (4.5) la verosimilitud del modelo será:

$$\begin{aligned}
\Pr(Y|X) &= \sum_{g} \Pr(g) \Pr(Y|g, X) \\
&= \sum_{g_1=1}^{k} \sum_{g_2=1}^{k} \dots \sum_{g_n=1}^{k} \prod_{i=1}^{n} p_i \ \Pr(Y_i|g, Y_1, Y_2, \dots, Y_{i-1}, X_i) \\
&= \sum_{g_1=1}^{k} \sum_{g_2=1}^{k} \dots \sum_{g_n=1}^{k} \prod_{i=1}^{n} p_i \left(\frac{e^{\omega_i(g_i)Y_i}}{1 + e^{\omega_i(g_i)}} \right)
\end{aligned} \tag{4.6}$$

con el producto sobre Y observables y las sumas sobre los posibles vectores g desconocidos.

Se define

$$H_i(g_1, g_2, \dots, g_{i-1}) = \begin{cases} p_i \left(\dfrac{e^{\omega_i(g_i)Y_i}}{1 + e^{\omega_i(g_i)}} \right), & si\ Y_i\ es\ observado \\ p_i, & si\ Y_i\ no\ es\ observado. \end{cases}$$

Con lo que (4.6) se puede escribir:

$$\Pr(Y|X) = \sum_{g_1=1}^{k} \sum_{g_2=1}^{k} \dots \sum_{g_n=1}^{k} \prod_{i=1}^{n} H_i(g_1, g_2, \dots, g_{i-1})$$

$$= \sum_{g_1=1}^{k} H_1(g_1) \sum_{g_2=1}^{k} H_2(g_1, g_2) \dots \sum_{g_n=1}^{k} \prod_{i=1}^{n} H_n(g_1, g_2, \dots, g_{n-1}). \tag{4.7}$$

Inferencia estadística

La realización de la estimación de los parámetros del modelo depende en gran parte de la calidad y el tipo de información con que se cuenta. Para el estudio de enfermedades se tienen bases de datos con diversos tipos de agrupamiento.

En el caso de observaciones sin parentesco o de las cuales no hay distinción ni información relativa a las relaciones parentales, los modelos indicados son el de dependencia en serie y el equitativamente predictivo ya que no toman en cuenta vínculos familiares.

Ahora bien, si se sabe información de los progenitores, la estructura de una familia nuclear, o un pedigrí, los modelos basados en dependencia biológica son los óptimos para obtener buenas aproximaciones.

Por lo anterior, es de suma importancia analizar el tipo de datos con los que se dispone, para así tener desde un inicio de forma clara y precisa el objetivo y método de la investigación.

La función de verosimilitud, dependerá entonces de que estructura familiar se estudia, que tipo de dependencia entre los individuos se sospecha existe y cuánta información se sabe sobre cada individuo. Después de plantear el modelo, se estiman los parámetros mediante alguna metodología estadística, y se determina la validez del modelo y la significancia que tienen las variables explicativas en la dependiente mediante, por ejemplo, los criterios y pruebas de hipótesis que se expusieron en el capítulo 3.

Ejemplo 4.1

Un ejemplo de la aplicación de la metodología expuesta en este trabajo en el ámbito del estudio de enfermedades genéticas, es el considerado por Bonney (1986).

En aquel trabajo el autor propone modelos para explicar una enfermedad metabólica llamada hiper alfalipoproteinemia familiar. La fuente de

información son 18 pedigrís publicados por Glueck et al. (1975) y la variable explicativa que se usó es el sexo de los individuos.

Bonney (1986) desarrolló cuatro modelos de los cuales uno no tiene ningún grado de dependencia entre observaciones y los otros tres son variantes del modelo clase A. Se presentarán los resultados obtenidos de dos de los modelos.

Se denota por Y=1 si el individuo presenta la enfermedad y cero si no la padece. La variable explicativa sexo será X=1 para mujeres y X=0 para hombres.

El primer modelo asume que las observaciones son independientes, es decir:

$$\omega_i = \alpha + \beta X_i, \quad i = 1,2,\dots,n.$$

El segundo modelo supone que las descendencias tienen el mismo grado de dependencia respecto del padre y madre asociados al individuo *i*, y también señala que hay algún tipo de relación entre la pareja y el individuo, es decir:

$$\omega_i(g_i) = \alpha_{g_i} + \gamma_C Z_C + \gamma_P (Z_P + Z_M) + \beta X_i, \quad i = 1,2,\dots,n.$$

Para comparar los modelos, Bonney (1986) usa el criterio de Akaike (1974) que señala que el mejor ajuste tiene el menor AIC definido como

$$AIC = -2\log(L) + 2(\sigma)$$

con
- L el valor de la verosimilitud evaluada en los estimadores de los parámetros y
- σ el número de parámetros estimados .

Los resultados por modelo son los siguientes:

	Modelo	Parámetros	-2LogL	AIC
1.	Ninguna dependencia	α, β	213.7	218
2.	Clase A	$\alpha, \gamma_C, \gamma_p = \gamma_M, \beta$	206.2	214*

El modelo que mejor ajusta es el segundo, del cual se obtuvieron los estimados siguientes:

Parámetro	Estimado	Error estándar
α	-0.98	0.262
γ_C	-0.79	0.5
$\gamma_p = \gamma_M$	-0.42	0.220
β	1.09	0.334

Dichos parámetros dan lugar a la ecuación estimada:

$$\omega = -0.98 - 0.79Z_C - 0.42Z_P - 0.42Z_M + 1.09X.$$

Con ω denotamos el *logit* estimado de presentar la enfermedad dados los valores de la pareja, los padres y el sexo del individuo. La relación mostrada entre los padres y la pareja es una correlación negativa, los valores tomados no son tan significativos como el de la variable explicativa, β, el cual sugiere que las mujeres tienen más posibilidades de desarrollar la enfermedad.

Conclusiones

El análisis de la información que brinda el entorno, investigaciones o simples tareas cotidianas siempre ha sido de interés, por lo que los esfuerzos por comprender y medir lo que nos rodea han rendido frutos a lo largo del tiempo.

Entre las herramientas que nos muestran como cuantificar y entender el medio que nos rodea, se encuentra la estadística y, con ella, el respaldo científico brindado por la teoría de probabilidad y matemáticas.

En la presente tesina se estudió una la metodología que permite conocer y, a su vez, estudiar y tratar rasgos genéticos. Se toman en cuenta los factores que intervienen en la segregación de los genes entre las generaciones de un pedigrí y cualquier factor o factores que se sospeche sean relevantes en la manifestación de la característica estudiada. Por ello, se brinda un enfoque integral al estudio de presentación de fenotipos en los descendientes.

La inferencia estadística recae sobre el uso de la regresión logística como eje de la propuesta de estudio. Este tipo de regresión permite explicar una variable dicotómica mediante variables de cualquier naturaleza y con la razón de momios, inherente a la regresión, se explica de forma práctica e ilustrativa la repercusión de cada variable explicativa en el modelo. En el capítulo 2 se muestra el amplio uso que tienen los modelos de regresión logística hoy en día.

Al unir el cálculo de la probabilidad de observar el rasgo estudiado en algún o algunos miembros de un árbol genealógico con el modelo de regresión logística se tienen distintos enfoques, tratados en el capítulo 4, para analizar los datos disponibles. Se toman en cuenta escenarios en que la información es escaza o faltante y se describen propuestas de modelos, así como recomendaciones con rigor estadístico para elegir el más acorde a lo estudiado. Se presenta la manera en que se infiere sobre lo obtenido y el pronunciamiento adecuado sobre los resultados obtenidos.

Con la presente metodología se cuenta con una forma de obtener un correcto y útil análisis con repercusión directa en la prevención y tratamiento de enfermedades hereditarias y para el conocimiento de los factores de riesgo potenciales que pueden provocar una patología.

Con lo anterior, se puede verificar que la estadística tiene alcances inimaginables en vastas áreas del conocimiento, siendo la epidemiología, la medicina y la genética algunos ejemplos. Entonces, el deber de un Actuario, aparte de crear e implementar modelos, es el interactuar y complementar lo estudiado por otras ciencias y especialistas, con la búsqueda de que al tener un equipo multidisciplinario, se conozca el comportamiento real de cualquier fenómeno que se pretenda estudiar.

Apéndice

Código correspondiente a la gráfica de la función logística para la figura 2

```
## Función logística  ##


x <- c(-8:8)

y <- 1/(1+(exp(-x)))

plot(x,y,xlab="z",ylab="f(z)",ylim=c(0,1.1),main="Función
logística",type="l",col=4)

abline(h=0,lty=9)

abline(h=1,lty=9)
```

Bibliografía

Agresti A. (2007). An introduction to Categorical Data Analysis. John Wiley & Sons, Inc.

Akaike, H. (1974). A new look al the statistical model identification. *IEEE Transactions of Automatic Control.* **AC-19**, 716-723.

Bonney, G. E. (1986). Regressive Logistic Models for Familial Disease and Other Binary Traits. *Biometrics* **42,** 611:625.

Cannings, C., Thompson E. A. y Skolnick, M. H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability* ***10,*** *26-61.*

Casella, G. y Berger, R. L. (2002). Statistical Inference. 2da edición. Thomson.

Elston, R. C. and Stewart, J. (1971) A general model for the genetic analysis of pedigree data. *Human Heredity* **21,** 523:542.

Glueck. C. J., Fallat. R. W., Millet. F., Gartside. P., Elston. R. C., y Go. R. C. P. (1975). Familial hyper-alpha-lipoproteinemia: Studies in eighteen kindreds. *Metabolism* **24,** 1243-1265.

Montgomery, D. C., Peck E. A. y Vining G. G. (2006) Introducción al Análisis de Regresión Lineal. 3era edición. Compañía editorial continental.

Pierce, B. A. (2009). Genética: Un enfoque conceptual. 2da edición. Editorial médica panamericana.