UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

FACULTAD DE CIENCIAS

Modelos de regresión aplicados a problemas de genética.

T E S I N A.

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A:

JIMENA PAOLA MERCADO RUIZ

DIRECTORA: DRA. ELIANE REGINA RODRIGUES

CIUDAD UNIVERSITARIA, CD. MX. 2017

Dedicatorias

A mis padres Verónica y Osbaldo, por siempre creer en mí, por brindarme su apoyo y cariño incondicional que me han motivado a alcanzar mis metas. Les agradezco todo su esfuerzo para darme una carrera, sin ustedes a mi lado no lo hubiera logrado. ¡Los quiero mucho!

A mi hermana Julieta, gracias por todos los momentos que pasamos juntas, por tu comprensión y tu ayuda cuando más lo necesitaba .

A mis tíos Rey y Jesús, y a prima Jess por todo el apoyo y el cariño que me han otorgado... muchas gracias por adoptarme.

A mis abuelitos por quererme y apoyarme siempre; y a todos mis tíos y primos muchas gracias por estar conmigo, ésto también se lo debo a ustedes.

Agradecimientos

A mi tutora, la Dra. Eliane Rodrigues, por haberme brindado la oportunidad de trabajar con ella, por su paciencia y apoyo en la realización de este trabajo.

A la Universidad Autónoma de México por la gran experiencia de cursar mi licenciatura en esta máxima casa de estudios.

A los sinodales encargados de revisar y corregir este trabajo, Dra. María del Pilar Alonso Reyes, Dra. María Asunción Begoña Fernández Fernández, Dra. Ana Meda Guardiola y al Act. Jaime Vázquez Alamilla.

A mis compañeros y amigos de la Facultad de Ciencias por todos los momentos que pasamos juntos, por las tardes de estudio y por cada una de las experiencias que compartimos.

Índice general

Introducción 1

1. Modelo básico de probabilidad para el análisis de datos de pedigrees **4**
1.1. Hipótesis 4
1.2. Notación 5
1.2.1. Notación utilizada para los datos 5
1.2.2. Definiciones Generales 5
1.3. Función de Verosimilitud 7
1.3.1. Función de verosimilitud de una generación de hermanos, dados los genotipos de sus padres ($\sigma(M_{X_i}) = s, \sigma(F_{X_i}) = t$) 7
1.3.2. Función de verosimilitud de una generación de cónyuges, dados los genotipos de los padres de su pareja ($\sigma(M_{X_i}) = s, \sigma(F_{X_i}) = t$) 9
1.3.3. Función de verosimilitud de la j-ésima generación de hermanos y sus respectivos cónyuges 10
1.3.4. Función de verosimilitud de un pedigree dado 11

2. Regresión lineal múltiple **14**
2.1. Modelo General 14
2.2. Estimación de los coeficientes β_j por mínimos cuadrados 16
2.3. Intervalo de confianza para los coeficientes β_j 24
2.4. Pruebas de hipótesis 26
2.4.1. Prueba de significancia de la regresión. 26

3. Modelo de regresión lineal múltiple para el análisis de mecanismos genéticos **31**
3.1. Notación 31
3.2. Probabilidad conjunta de los genotipos del pedigree 32
3.3. Distribución condicional de los residuales 33
3.4. Covarianzas 34
3.4.1. Covarianza conyugal. 34
3.4.2. Covarianza entre miembros del pedigree. 35
3.4.3. Covarianza dentro de un conjunto de hermanos. 35

Conclusiones 41

Referencias 42

Introducción

La mayoría de las enfermedades y rasgos humanos tienen un componente genético que puede ser heredado o influenciado por factores ambientales o de conducta.

En algunas ocasiones, efectivamente, la expresión anormal de uno o más genes, que fueron heredados por los progenitores a través de varias generaciones se manifiesta en enfermedades. Aunque éstas pueden ser tan inofensivas como el daltonismo, también lo pueden ser tan graves como ciertos tipos de cáncer.

De manera que, establecer la presencia de un mecanismo genético, es decir, la forma en que la transmisión de ciertos genotipos influye en rasgos observables específicos, es trascendental para identificar en los individuos el riesgo de desarrollar una enfermedad, incluso antes de que aparezcan los síntomas, y así evitar o retrasar sus complicaciones y repercusiones.

Incluso el análisis de la frecuencia de ciertos mecanismos genéticos que incrementan el riesgo de desarrollar enfermedades dentro de un árbol genealógico dado es fundamental, la razón de esto es que al heredarse el genotipo de progenitores a descendientes sus variaciones se conservan a lo largo de la línea familiar, de esta forma, se podría decir que cada familia comparte la variación de un mismo gen, la cual podría estar relacionada con la susceptibilidad a ciertos padecimientos.

Así pues, el objetivo de esta tesina es precisamente analizar determinados modelos estadísticos para el estudio de los mecanismos genéticos de un árbol genealógico dado.

De inicio, se plantean modelos de probabilidad, cuyo propósito es determinar la transmisión de información genética de un individuo a su descendiente y como es que esta información se expresa en los rasgos visibles de un individuo.

Posteriormente, se construye un modelo general de regresión lineal múltiple, asimismo se describen las características fundamentales de este tipo de modelos estadísticos.

Finalmente, se plantean determinados modelos estadísticos para los cuales se calcula la distribución de sus residuales. Es preciso señalar que estos modelos son distintos entre sí puesto que cada uno toma respectivamente en consideración la covarianza conyugal, la covarianza entre miembros del pedigree y la covarianza dentro de un conjunto de hermanos.

El conocimiento sobre las particularidades de la transmisión genética de un individuo a su descendiente y como es que esta información genética transmitida se expresa, está comenzando a modificar la forma como se practica la medicina, lo que hace pensar que en un futuro la atención médica será personalizada y estará enfocada a la predicción de ciertas variaciones genéticas que afectan nuestra salud.

Al abordar el tema de *Modelos de regresión compuestos para datos de familia* resulta necesario tener conocimiento de algunos conceptos de genética, por esta razón a continuación se incluyen nociones básicas de la materia.

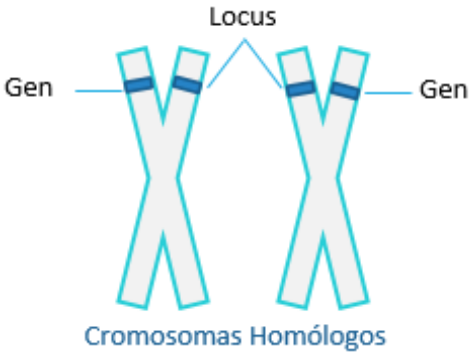
GENÉTICA

Ciencia encargada de estudiar la transmisión de la información que se hereda de generación en generación.

¿Cómo se hereda esta información?

Un gen es la unidad funcional básica de la herencia, en él se encuentra toda la información responsable de que algún rasgo en particular se exprese o no.

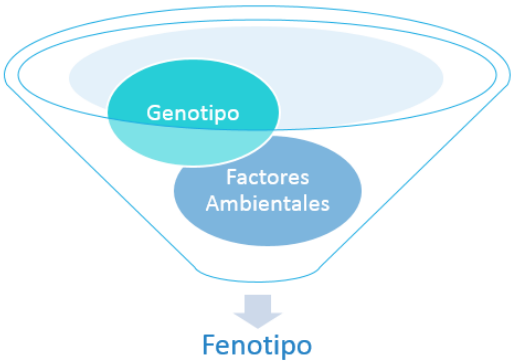
Los genes son segmentos de ácido desoxirribonucleico (ADN), molécula encargada de llevar la información genética para el desarrollo y funcionamiento los organismos vivos. Además, los genes se localizan en los cromosomas, en un lugar específico llamado locus, que en plural es loci.



Ahora, un gen puede tener dos o más variaciones. Estas variaciones son causadas por los alelos, que encuentran en la misma posición dentro de los cromosomas homólogos. Estos alelos son heredados por pares a los individuos, uno de ellos proviene de su madre y el otro de su padre.

Se llama genotipo al conjunto de genes de un individuo (incluyendo a sus alelos). Aunque éste no sea observable directamente, es de gran importancia, pues determina los rasgos observables de un individuo como, por ejemplo el color de ojos, a los cuales se les denomina fenotipo.

El genotipo es una característica que no se puede modificar, es decir, se mantiene constante a lo largo de la vida de un individuo. Mientras que el fenotipo, después de haber recibido la información genética, puede cambiar en respuesta a factores ambientales.



¿De qué sirve conocer esta información?

Los patrones hereditarios típicos se pueden reconocer, puesto que cada tipo de herencia (rasgos o enfermedades) produce árboles familiares únicos. Por lo cual, el análisis de árboles genealógicos, o pedigrees es sumamente importante para:

1. Establecer la presencia o ausencia de un mecanismo genético para la manifestación de un rasgo en particular o un conjunto de rasgos.

2. Explicar tal mecanismo, si está presente.

3. Clasificar los individuos por sus genotipos.

Capítulo 1

Modelo básico de probabilidad para el análisis de datos de pedigrees

Un modelo de probabilidad permite especificar la transmisión genética de un individuo a su descendiente y como es que esta información genética transmitida se expresa en lo que sí se puede observar (fenotipo).

En particular, el modelo básico de probabilidad presentado en este capítulo, se basa en los trabajos de Elston y Stewart[1971] y Cannings et, al. [1978].

1.1. Hipótesis

Para la construcción del modelo básico de probabilidad de este capítulo, se considerarán las hipótesis propuestas por Cannings et, al. [1978]:

i) A un individuo A del pedigree, se le asocia:
- $\sigma(A)$ - Genotipo.
- $\phi(A)$ - Fenotipo.
- $e(A)$ - Conjunto de factores ambientales.

ii) Para cada individuo, el fenotipo o rasgo observado depende del genotipo y de la respuesta a ciertos factores ambientales.

iii) Dado el genotipo $\sigma(A)$ del individuo A, su fenotipo $\phi(A)$ es independiente de los fenotipos y genotipos de cualquier otro conjunto de individuos.

iv) Dados los genotipos de la madre $\sigma(M_A)$ y del padre $\sigma(F_A)$ del individuo A, el genotipo de éste $\sigma(A)$ únicamente depende de $(\sigma(M_A), \sigma(F_A))$.

v) Dados el genotipo de la madre $\sigma(M_A)$ y el genotipo del padre $\sigma(F_A)$, los genotipos $\sigma(A_1), \sigma(A_2), ..., \sigma(A_r)$ de la descendencia $A_1, A_2, ..., A_r$ son mutuamente independientes.

1.2. Notación

1.2.1. Notación utilizada para los datos

Para identificar a cada miembro del pedigree, y debido a que únicamente se analizará el caso en el que no existen matrimonios consanguíneos, se utilizará la notación mencionada a continuación.

Existen 2 tipos de miembros:

- X : Individuo que está emparentado con alguien de la generación previa.
- Y : Individuo no relacionado con la generación previa, pero es pareja de un individuo que sí lo está.

Entonces, se definen como X_{i_0} , Y_{i_0} a los progenitores fundadores, es decir, los primeros individuos del i_0-ésimo pedigree.

De esta forma, $X_{i_0 i_1}$ es el i_1-ésimo hijo de estos progenitores fundadores y $Y_{i_0 i_1}$ su cónyuge. De igual manera $X_{i_0 i_1 i_2}$ es el i_2-ésimo hijo del i_1-ésimo hijo de los progenitores fundadores y $Y_{i_0 i_1 i_2}$ su cónyuge .

Por lo que, en general se tiene que, un individuo de la j-ésima generación del pedigree será de la forma $X_{i_0 i_1 ... i_{j-1} i_j}$, siendo el i_j-ésimo hijo del i_{j-1}-ésimo hijo del i_{j-2}-ésimo hijo del... del i_2-ésimo hijo del i_1-ésimo del i_0-ésimo individuo (progenitores fundadores). Mientras que su respectivo cónyuge, se denotará por $Y_{i_0 i_1 ... i_{j-1} i_j}$.

1.2.2. Definiciones Generales

Sea k el número de genotipos que causan variaciones en el fenotipo o rasgo que se desea observar. Entonces, se trata del mínimo número de genotipos distinguibles que deben existir en la población para explicar estas variaciones en un pedigree en particular. Así se tiene un u-ésimo genotipo con $u = 1, 2, ..., k$.

Entonces, tomando $x_1, x_2, ..., x_n$ los valores de x (del fenotipo) de una hermandad de n individuos, se define lo siguiente:

i) *Probabilidad de transición*, dada por:

$$\mathbb{P}(u|s,t) = \mathbb{P}\left[\sigma(A) = u|\sigma(M_A) = s, \sigma(F_A) = t\right] = p_{stu}.$$

Es decir, se trata de la probabilidad de que un individuo A tenga genotipo u dado que los genotipos de sus progenitores sean s y t. Con $u, s, t = 1, 2..., k$.

ii) *Función de penetrancia* dada por:

$$\mathbb{P}(x|u) = \mathbb{P}\left[\phi(A) \text{ observado sea } x|\sigma(A) = u, e(A)\right] = g_u(x).$$

Si el fenotipo de A no fue observado, entonces:

$$g_u(x) = 1 \qquad \text{para toda } u = 1, 2, ..., k.$$

En particular para este trabajo no se tomarán en cuenta los factores ambientales, de esta forma la probabilidad de que el rasgo x sea observado en el individuo A, dado el u-ésimo genotipo estará dada por:

$$g_u(x) = \mathbb{P}(u|\sigma(A) = u).$$

Ejemplos

i) Con la finalidad de entender mejor como se calculan las probabilidades de transición, se considera un modelo con un único locus y dos alelos (A, a):

Note que existen 3 genotipos dados por las combinaciones posibles entre los alelos:
$g_1 = AA$, $g_2 = Aa$ y $g_3 = aa$.

Entonces para la construcción de la *matriz de transición*, se va a denotar al genotipo de la madre por g_M y al genotipo del padre por g_F. De tal forma que para calcular cada entrada se necesitará calcular las siguientes probabilidades:

$$[\mathbb{P}(AA \mid g_M, g_F) \qquad \mathbb{P}(Aa \mid g_M, g_F) \qquad \mathbb{P}(aa \mid g_M, g_F)]$$

Obteniendo la siguiente *matriz de transición*:

		g_M		
		g_M=AA	g_M=Aa	g_M=aa
g_F	g_F=AA	[1 0 0]	[1/2 1/2 0]	[0 1 0]
	g_F=Aa	[1/2 1/2 0]	[1/4 1/2 1/4]	[0 1/2 1/2]
	g_F=aa	[0 1 0]	[0 1/2 1/2]	[0 0 1]

Así, por ejemplo para la primer entrada de la segunda fila, se tienen dados los genotipos $g_M = AA$ y $g_F = Aa$, que forman las combinaciones $\{AA, Aa, AA, Aa\}$.

Dadas estas combinaciones, se calcula lo siguiente:

1)$\mathbb{P}(AA \mid AA, Aa) = \frac{2}{4} = \frac{1}{2}$.

2)$\mathbb{P}(Aa \mid AA, Aa) = \frac{2}{4} = \frac{1}{2}$.

3)$\mathbb{P}(aa \mid AA, Aa) = 0$.

Que precisamente fue lo que se obtuvo: $\left[\frac{1}{2} \qquad \frac{1}{2} \qquad 0\right]$.

ii) Para comprender mejor el comportamiento de la función de penetrancia, se consideran los grupos sanguíneos existentes, que se forman a partir de 3 alelos (A, B, O):

g_M	g_F	Combinaciones	Resultado
A	A	AA	A
A	B	AB	AB
A	O	AO	A
B	A	AB	AB
B	B	BB	B
B	O	BO	B
O	O	OO	O

En la tabla anterior se puede observar que:

1)A y B son dominantes respecto a O.

2)O es recesivo respecto a A y a B.

3)A y B son codominantes.

Dado lo anterior, cada función de probabilidad va a ser asociada con un genotipo, debido a que, para la herencia cuantitativa cada genotipo está asociado a un rango de valores fenotípicos, aunque la variación en cada genotipo se debe a influencias ambientales.

1.3. Función de Verosimilitud

Con la finalidad de que la construcción del caso general de la función de verosimilitud sea más comprensible, se considerarán los siguientes casos previos.

1.3.1. Función de verosimilitud de una generación de hermanos, dados los genotipos de sus padres $(\sigma(M_{X_i}) = s, \sigma(F_{X_i}) = t)$

Para el individuo X_i, se tiene que la probabilidad de que muestre el fenotipo x_i, dados los genotipos de sus padres, se expresa como:

$\mathbb{P}\left[X_i \text{ muestre el rasgo } x_i \mid \sigma(M_{x_i}) = s, \sigma(F_{x_i}) = t\right]$. Que en adelante se denotará por $\mathbb{P}\left(x_i \middle| s,t\right)$.

De la misma forma, para una generación de n hermanos, dicha probabilidad está dada por $\mathbb{P}\left(x_1, x_2, ..., x_n \middle| s,t\right)$.

Por la definición de función de densidad marginal del vector discreto (X,Y), se tiene que:

$$\mathbb{P}(x_i) = \sum_j \mathbb{P}(x_i, y_j).$$

Dada la partición $\bigcup_{i=1}^{k} \{\sigma(X) = i\}$ formada por los $u = 1, ..., k$ genotipos posibles, se puede escribir:

$$\mathbb{P}\left(x_1, ..., x_n \middle| s,t\right) = \sum_{u_1=1}^{k} \sum_{u_2=1}^{k} ... \sum_{u_n=1}^{k} \mathbb{P}\left(x_1, x_2, ..., x_n, u_1, u_2, ..., u_n \mid s,t\right). \tag{2.3.1}$$

Por el supuesto iv) dado en la sección 2.1, se tiene que: dados los genotipos de la madre $\sigma(M_{X_i}) = s$ y de el padre $\sigma(F_{X_i}) = t$ del individuo X_i, el genotipo de éste únicamente depende de (s,t).

Por lo que, la expresión (2.3.1) es equivalente a:

$$\begin{aligned}
&= \sum_{u_1=1}^{k} \sum_{u_2=1}^{k} ... \sum_{u_n=1}^{k} \mathbb{P}\left(x_1, u_1 \mid s,t\right) \mathbb{P}\left(x_2, u_2 \mid s,t\right) ... \mathbb{P}\left(x_n, u_n \mid s,t\right) \\
&= \sum_{u_1=1}^{k} \mathbb{P}\left(x_1, u_1 \mid s,t\right) \sum_{u_2=1}^{k} \mathbb{P}\left(x_2, u_2 \mid s,t\right) ..... \sum_{u_n=1}^{k} \mathbb{P}\left(x_n, u_n \mid s,t\right) \\
&= \prod_{i=1}^{n} \left[\sum_{u=1}^{k} \mathbb{P}\left(x_i, u \mid s,t\right) \right] \\
&= \prod_{i=1}^{n} \left[\sum_{u=1}^{k} \mathbb{P}\left(x_i \mid u,s,t\right) \mathbb{P}\left(u \mid s,t\right) \right].
\end{aligned} \tag{2.3.2}$$

Ahora, por el supuesto iii) dado en igualmente en la sección 2.1, se tiene que: dado el genotipo $\sigma(X_i)$, el fenotipo del individuo X_i es independiente de los fenotipos y genotipos de cualquier otro conjunto de individuos. De esta forma, si se sabe que $\sigma(X_i) = u_i$, no se requieren más de los genotipos de los padres (s,t) para calcular la probabilidad de que X_i tenga el fenotipo x_i.

En consecuencia, (2.3.2) es igual a:

$$= \prod_{i=1}^{n} \left[\sum_{u=1}^{k} \underbrace{\mathbb{P}\left(x_i \mid u\right)}_{g_u(x_i)} \underbrace{\mathbb{P}\left(u \mid s,t\right)}_{p_{stu}} \right].$$

Por lo que,

$$\mathbb{P}\left(x_1, ..., x_n | \, s, t\right) = \prod_{i=1}^{n} \left[\sum_{u=1}^{k} g_u(x_i) p_{stu} \right]. \tag{2.3.3}$$

1.3.2. Función de verosimilitud de una generación de cónyuges, dados los genotipos de los padres de su pareja $(\sigma(M_{X_i}) = s, \sigma(F_{X_i}) = t)$

Para el cónyuge Y_i del $i-$ ésimo miembro, usando la notación definida en la sección 2.3.1, se tiene que la probabilidad de que muestre el fenotipo y_i, dados los genotipos de los padres de su respectiva pareja, se expresa como, $\mathbb{P}\left(y_i \mid s, t\right)$.

Ahora, como el individuo Y_i no está emparentado con los demás individuos del pedigree, la probabilidad de observar el rasgo y_i no depende de los genotipos s, t.

Entonces, para n cónyuges de una generación, $\mathbb{P}\left(y_1, y_2..., y_n \mid s, t\right) = \mathbb{P}\left(y_1, ..., y_n\right)$.

Igualmente, para la partición $\bigcup_{i=1}^{k} \{\sigma(Y) = i\}$ formada por los $v = 1, ..., k$ genotipos posibles y usando la definición de probabilidad marginal, así como la ley de probabilidad total, se tiene que:

$$\begin{aligned}
\mathbb{P}\left(y_1, y_2, ..., y_n\right) &= \sum_{v_1=1}^{k} \sum_{v_2=1}^{k} ... \sum_{v_n=1}^{k} \mathbb{P}\left(y_1, y_2, ..., y_n, v_1, v_2, ..., v_n\right) \\
&= \sum_{v_1=1}^{k} \sum_{v_2=1}^{k} ... \sum_{v_n=1}^{k} \mathbb{P}\left(y_1 \mid v_1\right) \mathbb{P}\left(v_1\right) \mathbb{P}\left(y_2 \mid v_2\right) \mathbb{P}\left(v_2\right) .....\mathbb{P}\left(y_n \mid v_n\right) \mathbb{P}\left(v_n\right) \\
&= \sum_{v_1=1}^{k} \mathbb{P}\left(y_1 \mid v_1\right) \mathbb{P}\left(v_1\right) \sum_{v_2=1}^{k} \mathbb{P}\left(y_2 \mid v_2\right) \mathbb{P}\left(v_2\right) ..... \sum_{v_n=1}^{k} \mathbb{P}\left(y_n \mid v_n\right) \mathbb{P}\left(v_n\right) \\
&= \prod_{i=1}^{n} \left[\sum_{v=1}^{k} \underbrace{\mathbb{P}\left(y_i \mid u\right)}_{g_v(y_i)} \mathbb{P}\left(v\right) \right].
\end{aligned}$$

Se define a Ψ_v como la frecuencia genotípica, es decir, la proporción de individuos en la población que tienen el v-ésimo genotipo. Entonces $\Psi_v = \mathbb{P}\left(v\right)$.

Finalmente,

$$\mathbb{P}\left(y_1, y_2, ..., y_n\right) = \prod_{i=1}^{n} \left[\sum_{v=1}^{k} g_v(y_i) \Psi_v \right] \tag{2.3.4}$$

1.3.3. Función de verosimilitud de la j-ésima generación de hermanos y sus respectivos cónyuges

En este caso, la probabilidad conjunta de que los individuos X_i, Y_i presenten los rasgos x_i, y_i respectivamente, se expresa como $\mathbb{P}(x_i, y_i)$.

Entonces para n hermanos de una generación y sus cónyuges respectivos, se tendría,
$\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n)$.

De igual forma que los anteriores casos, usando probabilidad marginal y la ley de probabilidad total:

$$\begin{aligned}\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n) &= \sum_{s=1}^{k}\sum_{t=1}^{k}\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n, s, t)\\ &= \sum_{s=1}^{k}\sum_{t=1}^{k}\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n \mid s, t)\,\mathbb{P}(s, t)\\ &= \sum_{s=1}^{k}\sum_{t=1}^{k}\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n \mid s, t)\,\mathbb{P}(s)\,\mathbb{P}(t)\\ &= \sum_{s=1}^{k}\sum_{t=1}^{k}\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n \mid s, t)\,\Psi_s\Psi_t.\end{aligned}$$

Ahora, se observa que:

$$\begin{aligned}&\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n \mid s, t) =\\ &= \sum_{u_1=1}^{k}...\sum_{u_n=1}^{k}\sum_{v_1=1}^{k}...\sum_{v_n=1}^{k}\mathbb{P}(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n, u_1, ..., u_n, v_1, ..., v_n \mid s, t) =\\ &= \sum_{u_1=1}^{k}...\sum_{u_n=1}^{k}\sum_{v_1=1}^{k}...\sum_{v_n=1}^{k}\mathbb{P}(x_1, u_1 \mid s, t)...\mathbb{P}(x_n, u_n \mid s, t)\,\mathbb{P}(y_1, v_1 \mid s, t)...\mathbb{P}(y_n, v_n \mid s, t)\\ &= \sum_{u_1=1}^{k}...\sum_{u_n=1}^{k}\sum_{v_1=1}^{k}...\sum_{v_n=1}^{k}\mathbb{P}(x_1 \mid u_1)\,\mathbb{P}(u_1 \mid s, t)...\mathbb{P}(x_n \mid u_n)\,\mathbb{P}(u_n \mid s, t)\\ &\qquad\qquad \mathbb{P}(y_1 \mid v_1)\,\mathbb{P}(v_1 \mid s, t)...\mathbb{P}(y_n \mid v_n)\,\mathbb{P}(v_n \mid s, t)\\ &= \sum_{u_1=1}^{k}\mathbb{P}(x_1 \mid u_1)\,\mathbb{P}(u_1 \mid s, t)...\sum_{u_n=1}^{k}\mathbb{P}(x_n \mid u_n)\,\mathbb{P}(u_n \mid s, t)\\ &\qquad\qquad \sum_{v_1=1}^{k}\mathbb{P}(y_1 \mid v_1)\,\mathbb{P}(v_1)...\sum_{v_n=1}^{k}\mathbb{P}(y_n \mid v_n)\,\mathbb{P}(v_n)\\ &= \left[\prod_{i=1}^{n}\left(\sum_{u=1}^{k}\underbrace{\mathbb{P}(x_i \mid u)}_{g_u(x_i)}\underbrace{\mathbb{P}(u \mid s, t)}_{p_{stu}}\right)\right]\left[\prod_{i=1}^{n}\left(\sum_{v=1}^{k}\underbrace{\mathbb{P}(y_i \mid v)}_{g_v(y_i)}\underbrace{\mathbb{P}(v)}_{\Psi_v}\right)\right].\end{aligned}$$

Por lo tanto:

$$\mathbb{P}\left(x_1, x_2, ..., x_n, y_1, y_2, ..., y_n\right) =$$

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

Se puede ver que esta expresión está en función de los genotipos s y t, y no sólo eso, además estos genotipos corresponden a una expresión similar para los genotipos u y v, pero de la generación anterior.

Entonces, para poder expresar esta relación entre generaciones, se va a reescribir, obteniendo que, la verosimilitud de observar a la j-ésima generación de los miembros la hermandad y sus cónyuges está dada por:

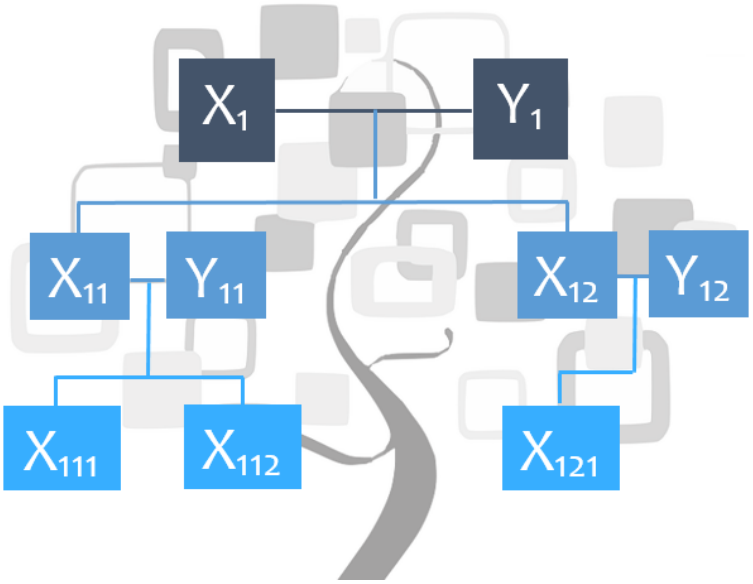
$$\mathbf{\Gamma_j} = \prod_{i_j=1}\sum_{s_j=1}^{k} p_{s_{j-1}t_{j-1}s_j} g_{s_j}(x_{i_0 i_1 \ldots i_j}) \sum_{t_j=1}^{k} \Psi_{t_j} g_{t_j}(y_{i_0 i_1 \ldots i_j}). \tag{2.3.5}$$

1.3.4. Función de verosimilitud de un pedigree dado

Antes de presentar el caso general de la función de verosimilitud, se verá el siguiente ejemplo:

Usando la notación descrita en la sección 2.2.1, el genotipo de cada miembro X_{ij}, Y_{ij}, estará dado por las letras u_{ij} y v_{ij} respectivamente; a excepción de los genotipos de los progenitores originales denotados como $\sigma(M_{X_{ij}}) = s$ y $\sigma(F_{X_{ij}}) = t$. Además, los fenotipos de los hermanos y sus cónyuges estarán dados por $\phi(X_{ij}) = x_{ij}$ y $\phi(Y_{ij}) = y_{ij}$.

Suponga que se tiene el siguiente pedigree:



Para obtener la función de verosimilitud de los miembros de este pedigree, se necesita $\mathbb{P}(x_1, y_1, x_{11}, y_{11}, x_{12}, y_{12}, x_{111}, x_{112}, x_{121})$.

Primero usando probabilidad marginal, se puede escribir:

$$\begin{aligned}&\mathbb{P}(x_1, y_1, x_{11}, y_{11}, x_{12}, y_{12}, x_{111}, x_{112}, x_{121})\\ =&\sum_{s=1}^{k}\sum_{t=1}^{k}\sum_{u_{11}=1}^{k}\sum_{v_{11}=1}^{k}\sum_{u_{12}=1}^{k}\sum_{v_{12}=1}^{k}\sum_{u_{111}=1}^{k}\sum_{u_{112}=1}^{k}\sum_{u_{121}=1}^{k}\\ &\mathbb{P}(x_1, y_1, x_{11}, y_{11}, x_{12}, y_{12}, x_{111}, x_{112}, x_{121}, s, t, u_{11}, v_{11}, u_{12}, v_{12}, u_{111}, u_{112}, u_{121}).\end{aligned}$$

Además, por propiedades de probabilidad condicional, dada $A_1, A_2, ..., A_n$ una colección de eventos aleatorios tal que $\mathbb{P}\left[\bigcap_{i=2}^{i=n} A_i\right] > 0$, entonces,

$$\begin{aligned}\mathbb{P}(A_1, A_2, ..., A_n) &= \mathbb{P}(A_1 \mid A_2, A_2, ..., A_n)\mathbb{P}(A_2, A_3, ..., A_n)\\ &= \mathbb{P}(A_1 \mid A_2, A_2, ..., A_n)\left[\mathbb{P}(A_2 \mid A_3, ..., A_n)\mathbb{P}(A_3, ..., A_n)\right]\\ &= \mathbb{P}(A_1 \mid A_2, A_2, ..., A_n)\mathbb{P}(A_2 \mid A_3, ..., A_n).....\mathbb{P}(A_{n-1} \mid A_n)\mathbb{P}(A_n).\end{aligned}$$

De esta forma, queda que:

$$\begin{aligned}&\mathbb{P}(x_1, y_1, x_{11}, y_{11}, x_{12}, y_{12}, x_{111}, x_{112}, x_{121})\\ =&\sum_{s=1}^{k}\sum_{t=1}^{k}\sum_{u_{11}=1}^{k}\sum_{v_{11}=1}^{k}\sum_{u_{12}=1}^{k}\sum_{v_{12}=1}^{k}\sum_{u_{111}=1}^{k}\sum_{u_{112}=1}^{k}\sum_{u_{121}=1}^{k}\\ &\{\mathbb{P}(x_1 \mid s)\mathbb{P}(y_1 \mid t)\mathbb{P}(x_{11} \mid u_{11})\mathbb{P}(y_{11} \mid v_{11})\mathbb{P}(x_{12} \mid u_{12})\mathbb{P}(y_{12} \mid v_{12})\\ &\mathbb{P}(x_{111} \mid u_{111})\mathbb{P}(x_{112} \mid u_{112})\mathbb{P}(x_{121} \mid u_{121})\mathbb{P}(u_1 \mid s, t)\mathbb{P}(v_1)\mathbb{P}(u_{11} \mid s, t)\mathbb{P}(v_{11})\\ &\mathbb{P}(u_{12} \mid s, t)\mathbb{P}(v_{12})\mathbb{P}(u_{111} \mid s, t)\mathbb{P}(u_{112} \mid s, t)\mathbb{P}(u_{121} \mid s, t)\mathbb{P}(s, t)\}.\end{aligned}$$

Reagrupando,

$$\begin{aligned}=&\sum_{s=1}^{k}\sum_{t=1}^{k}\Bigg\{\sum_{s=1}^{k}\mathbb{P}(x_1 \mid s)\sum_{t=1}^{k}\mathbb{P}(y_1 \mid t)\\ &\Bigg[\sum_{u_{11}=1}^{k}\mathbb{P}(x_{11} \mid u_1)\mathbb{P}(u_{11} \mid s, t)\sum_{v_{11}=1}^{k}\mathbb{P}(y_{11} \mid v_{11})\mathbb{P}(v_{11})\\ &\sum_{u_{12}=1}^{k}\mathbb{P}(x_{12} \mid u_1)\mathbb{P}(u_{12} \mid s, t)\sum_{v_{12}=1}^{k}\mathbb{P}(y_{12} \mid v_{12})\mathbb{P}(v_{12})\\ &\Bigg(\sum_{u_{111}=1}^{k}\mathbb{P}(x_{111} \mid u_{111})\mathbb{P}(u_{111} \mid s, t)\sum_{u_{112}=1}^{k}\mathbb{P}(x_{112} \mid u_{112})\mathbb{P}(u_{112} \mid s, t)\\ &\sum_{u_{121}=1}^{k}\mathbb{P}(x_{121} \mid u_{121})\mathbb{P}(u_{121} \mid s, t)\Bigg)\Bigg]\mathbb{P}(s)\mathbb{P}(t)\Bigg\}.\end{aligned}$$

Recordando lo siguiente:

$i)\mathbb{P}(u \mid s,t) = p_{stu}$ es la probabilidad de que un individuo tenga genotipo u, dado que los genotipos de sus padres son s y t.

$ii)\mathbb{P}(x_i \mid u) = g_u(x_i)$ es la probabilidad de que se observe el fenotipo x_i dado que el individuo tiene genotipo u.

$iii)\mathbb{P}(m) = \Psi_m$ es la frecuencia genotípica (del m-ésimo genotipo).

Utilizando esto, se obtiene:

$$\mathbb{P}\left(x_1, y_1, x_{11}, y_{11}, x_{12}, y_{12}, x_{111}, x_{112}, x_{121}\right)$$

$$= \left\{ \sum_{s=1}^{k}\sum_{t=1}^{k} g_s(x_1)g_t(y_1) \right.$$
$$\left[\sum_{u_{11}=1}^{k} g_{u_{11}}(x_{11})p_{stu_{11}} \sum_{v_{11}=1}^{k} g_{v_{11}}(y_{11})\Psi_{v_{11}} \sum_{u_{12}=1}^{k} g_{u_{12}}(x_{12})p_{stu_{12}} \sum_{v_{12}=1}^{k} g_{v_{12}}(y_{12})\Psi_{v_{12}} \right.$$
$$\left. \left. \left(\sum_{u_{111}=1}^{k} g_{u_{111}}(x_{111})p_{u_{11}v_{11}u_{111}} \sum_{u_{112}=1}^{k} g_{u_{112}}(x_{112})p_{u_{11}v_{11}u_{112}} \sum_{u_{121}=1}^{k} g_{u_{121}}(x_{121})p_{u_{12}v_{12}u_{121}} \right) \right] \Psi_s\Psi_t \right\}.$$

Ahora, para que sea más comprensible la expresión anterior se reescribe de la siguiente manera:

$$\mathbb{P}\left(x_1, y_1, x_{11}, y_{11}, x_{12}, y_{12}, x_{111}, x_{112}, x_{121}\right) =$$

$$\sum_{s=1}^{k}\sum_{t=1}^{k} \left\{ g_s(x_1)g_t(y_1) \left[\prod_{i=1}^{2} \sum_{u_{1i}=1}^{k} g_{u_{1i}}(x_{1i})p_{stu_{1i}} \sum_{v_{1i}=1}^{k} g_{v_{1i}}(y_{1i})\Psi_{v_{1i}} \right.\right.$$
$$\left.\left. \left(\prod_{i,j=1}^{2} \sum_{u_{1ij}=1}^{k} g_{u_{1ij}}(x_{1ij})p_{u_{1i}v_{1i}u_{1ij}} \right) \right] \Psi_s\Psi_t \right\}.$$

Además, recordando por la expresión (2.3.5), que la verosimilitud de la j-ésima generación de hermanos y cónyuges está dada por Γ_j. Finalmente utilizando esta notación, se tiene que:

$$\mathbb{P}\left(x_1, y_1, x_{11}, y_{11}, x_{12}, y_{12}, x_{111}, x_{112}, x_{121}\right) = \Gamma_0(\Gamma_1(\Gamma_2)).$$

Donde:

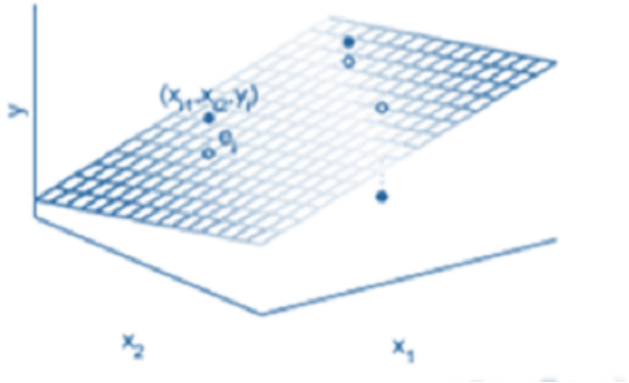
- Γ_0 corresponde a la primera generación del pedigree, es decir: x_1, y_1.
- Γ_1 corresponde a la segunda generación del pedigree, es decir: $x_{11}, y_{11}, x_{12}, y_{12}$
- Γ_2 corresponde a la tercera generación del pedigree, es decir: $x_{111}, x_{112}, x_{121}$

Por lo tanto, en general la verosimilitud de un pedigree dado puede ser expresada por:

$$\boldsymbol{\Gamma_0(\Gamma_1(\Gamma_2(\Gamma_3(...)))))}. \qquad \textbf{(2.3.6)}$$

Capítulo 2

Regresión lineal múltiple



En un modelo de regresión lineal simple, la medición de una sola variable respuesta está relacionada con una única variable explicativa. Mientras que, en la regresión lineal múltiple se pueden utilizar más de una variable explicativa, permitiendo emplear más información en la construcción del modelo y así realizar estimaciones más precisas.

2.1. Modelo General

Si existe más de una variable explicativa, por ejemplo $X_1, X_2, ..., X_p$, asociadas a una variable de respuesta Y, el modelo general estará dado por:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon,$
donde β_j son los coeficientes parciales de regresión, $j = 0, 1, 2, ..., p$.

Cabe mencionar que la notación utilizada en el presente capítulo será la usual, por lo cual ésta será diferente a la presentada en los demás capítulos.

De esta forma, suponiendo que se tienen $n > p$ observaciones disponibles, denotadas por $y_1, y_2, ..., y_p$ y sean $x_{i1}, x_{i2}, ..., x_{ip}$ los valores de las variables explicativas asociadas a la observación y_i. Entonces se tendrán los siguientes datos para el modelo:

Observación (i)	Variables de Respuesta (y_i)	Variables explicativas (x_{ij})			
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
3	y_3	x_{31}	x_{32}	...	x_{3p}
.	.	.	.		.
.	.	.	.	...	.
.	.	.	.		.
n	y_n	x_{n1}	x_{n2}	...	x_{np}

Por lo que, para la i-ésima observación el modelo estará dado por:

$$\mathbf{y_i} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x_{i1}} + \beta_\mathbf{2}\mathbf{x_{i2}} + ... + \beta_\mathbf{p}\mathbf{x_{ip}} + \epsilon_\mathbf{i}.$$

Donde:

- y_i es la i-ésima variable dependiente o de respuesta.
- x_{ij}, j=1,2,...,p representa a la j-ésima variable explicativa de la i-ésima observación.
- β_0 intersección (valor cuando todas las variables independientes son cero).
- β_j, j=1,2,...,p representa los correspondientes p coeficientes parciales de regresión.
- ϵ_i es el i-ésimo error aleatorio.

Para que el modelo quede escrito de una forma más compacta, se puede hacer lo siguiente:

1)Acomodar las variables de respuesta en un vector de dimensión n, es decir,

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}.$$

2)Acomodar todas las variables explicativas en una matriz de dimensión $n \times (p+1)$, es decir,

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2p} \\ & & ... & & & & \\ & & ... & & & & \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{pmatrix}.$$

Cabe destacar que la columna de "unos" corresponde a β_0.

3)Acomodar todos los coeficientes en un vector de dimensión p+1, es decir,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}.$$

4)Acomodar todos los errores aleatorios en un vector de dimensión n, es decir,

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}.$$

Finalmente, el modelo general estará dado por la siguiente expresión:

$$\boxed{Y = X\beta + \epsilon.} \tag{3.1.1}$$

Observación 3.1.1

En los modelos de regresión lineal, el coeficiente β_j de la variable explicativa X_j o **coeficiente parcial de regresión**, se interpreta como el cambio en Y, asociado a un cambio de una unidad en X_j, cuando las demás variables explicativas se mantienen constantes.

2.2. Estimación de los coeficientes β_j por mínimos cuadrados

Las hipótesis necesarias para la estimación de los coeficientes del presente capítulo están basadas en lo propuesto por Montgomery et al. [2006]. Entonces, suponga que los errores aleatorios cumplen lo siguiente:

i) $\mathbb{E}(\epsilon_i) = 0$
ii) $\mathbb{V}ar(\epsilon_i) = \sigma^2 > 0$
iii) $\mathbb{C}ov(\epsilon_i, \epsilon_j) = 0$ para toda $i \neq j$ es decir, están no correlacionados.

Considere la siguiente definición:

Definición 3.2.1

La función de mínimos cuadrados para un modelo de regresión con coeficientes $\beta = (\beta_0, \beta_1, ..., \beta_p)$ está definida por: $S(\beta) = \sum_{i=1}^{n} \epsilon_i^2$.

Observación 3.2.1

Note que, si ϵ' la transpuesta de la matriz ϵ, entonces,

$$\epsilon'\epsilon = \begin{pmatrix} \epsilon_1 & \epsilon_2 & . & . & . & \epsilon_n \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix} = \epsilon_1^2 + \epsilon_2^2 + ... + \epsilon_n^2 = \sum_{i=1}^{n} \epsilon_i^2.$$

Por lo tanto, la función de mínimos cuadrados es equivalente a: $S(\beta) = \epsilon'\epsilon$.

Lema 3.2.1

La función de mínimos cuadrados puede ser vista como:

$$S(\beta) = Y'Y - 2\beta'X'Y + \beta X'X\beta. \tag{3.2.1}$$

Demostración

Recordando que el modelo está dado por la expresión $Y = X\beta + \epsilon$.

Entonces, se puede escribir, $\epsilon = Y - X\beta$.

De esta forma,

$$S(\beta) = \epsilon'\epsilon = (Y - X\beta)'(Y - X\beta). \tag{3.2.2}$$

Por propiedades de matrices transpuestas, se sabe que $(A + B)' = A' + B'$ y que $(AB)' = B'A'$. Por lo que, la expresión (3.2.2) se transforma en:

$$S(\beta) = (Y' - (X\beta)')(Y - X\beta) = (Y' - \beta'X')(Y - X\beta)$$

$$= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta. \tag{3.2.3}$$

Ahora, para ver que $\beta'X'Y$ es una matriz de dimensión 1×1 primero se calculará $X'Y$:

$$X'Y = \begin{pmatrix} 1 & 1 & . & . & . & 1 \\ x_{11} & x_{21} & . & . & . & x_{n1} \\ & ... & & & & \\ & ... & & & & \\ x_{1p} & x_{2p} & . & . & . & x_{np} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} y_1 + y_2 + ... + y_n \\ x_{11}y_1 + x_{21}y_2 + ... + x_{n1}y_n \\ . \\ . \\ x_{1p}y_1 + x_{2p}y_2 + ... + x_{np}y_n \end{pmatrix}.$$

Entonces, queda que: $$X'Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip}y_i \end{pmatrix}.$$

Multiplicando por β',

$$\beta'X'Y = \begin{pmatrix} \beta_0, \beta_1, & . & . & \beta_p \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip}y_i \end{pmatrix} = \beta_0 \sum_{i=1}^{n} y_i + ... + \beta_p \sum_{i=1}^{n} x_{ip}y_i. \tag{3.2.4}$$

De esta forma queda demostrado que $\beta'X'Y$ es una matriz de dimensión 1×1, por lo que su transupuesta será la misma matriz de dimensión 1×1, es decir,

$$\beta'X'Y = (\beta'X'Y)'.$$

Además, se tiene que $(\beta'X'Y)' = Y'(\beta'X')' = Y'X\beta$.

Sustituyendo en (3.2.3):

$$\begin{aligned} S(\beta) &= Y'Y - \beta'X'Y - (\beta'X'Y)' + \beta X'X\beta \\ &= Y'Y - \beta'X'Y - \beta'X'Y + \beta X'X\beta. \end{aligned}$$

Por consiguiente, la función de mínimos cuadrados queda como:

$$S(\beta) = Y'Y - 2\beta'X'Y + \beta X'X\beta$$

Proposición 3.2.2

El valor de β que minimiza la función de mínimos cuadrados $S(\beta)$ es,

$$\beta = (X'X)^{-1}X'Y. \tag{3.2.5}$$

Demostración

Primero, se va a calcular la derivada de $S(\beta)$, para ello con las propiedades de la derivada:

$\frac{\partial S(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}(Y'Y - 2\beta'X'Y + \beta X'X\beta) = \frac{\partial}{\partial \beta}(Y'Y) - 2\frac{\partial}{\partial \beta}(\beta'X'Y) + \frac{\partial}{\partial \beta}(\beta X'X\beta)$

En los incisos $i), ii)$ y $iii)$, se va a desarrollar cada término de la suma.

i) $\frac{\partial}{\partial\beta}(Y'Y) = \frac{\partial}{\partial\beta}\begin{pmatrix} y_1 & y_2 & . & . & . & y_n \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \frac{\partial}{\partial\beta}(y_1^2 + y_2^2 + ... + y_n^2) = 0$

ii) $\frac{\partial}{\partial\beta}(\beta' X'Y)$
Por la expresión (3.2.4), se tiene que:

$$\beta' X'Y = \beta_0 \sum_{i=1}^{n} y_i + ... + \beta_p \sum_{i=1}^{n} x_{ip} y_i.$$

Además, sabiendo que si $u = f(x)$ es una función de $x_1, x_2, ..., x_n$, entonces

$$\frac{\partial u}{\partial x} = \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \frac{\partial u}{\partial x_2} \\ . \\ . \\ . \\ \frac{\partial u}{\partial x_n} \end{pmatrix}.$$

Por lo que,

$$\frac{\partial}{\partial\beta}(\beta' X'Y) = \begin{pmatrix} \frac{\partial}{\partial\beta_0}(\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_{i1} y_i) + ... + \beta_p \sum_{i=1}^{n} x_{ip} y_i) \\ \frac{\partial}{\partial\beta_1}(\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_{i1} y_i) + ... + \beta_p \sum_{i=1}^{n} x_{ip} y_i) \\ . \\ . \\ . \\ \frac{\partial}{\partial\beta_p}(\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_{i1} y_i) + ... + \beta_p \sum_{i=1}^{n} x_{ip} y_i) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ . \\ . \\ \sum_{i=1}^{n} x_{ip} y_i \end{pmatrix}$$

$$= X'Y.$$

De esta forma, es válido que $-2\frac{\partial}{\partial\beta}(\beta' X'Y) = -2X'Y$.

iii) $\frac{\partial}{\partial \beta}(\beta X'X\beta)$

Primero, calculando $X'X$,

$$X'X = \begin{pmatrix} 1 & 1 & . & . & . & 1 \\ x_{11} & x_{21} & . & . & . & x_{n1} \\ & ... & & & & \\ x_{1p} & x_{2p} & . & . & . & x_{np} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & . & . & . & x_{1p} \\ 1 & x_{21} & . & . & . & x_{2p} \\ & ... & & & & \\ 1 & x_{n1} & . & . & . & x_{np} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_{i1} & . & . & . & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & . & . & . & \sum_{i=1}^{n} x_{i1}x_{ip} \\ & & & ... & & \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{i1}x_{ip} & . & . & . & \sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}.$$

Multiplicando por β,

$$X'X\beta = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & . & . & . & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & . & . & . & \sum_{i=1}^{n} x_{i1}x_{ip} \\ & & & ... & & \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{i1}x_{ip} & . & . & . & \sum_{i=1}^{n} x_{ip}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}$$

$$= \begin{pmatrix} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{i1} + .. + \beta_p \sum_{i=1}^{n} x_{ip} \\ \beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + ... + \beta_p \sum_{i=1}^{n} x_{i1}x_{ip} \\ ... \\ \beta_0 \sum_{i=1}^{n} x_{ip} + \beta_1 \sum_{i=1}^{n} x_{i1}x_{ip} + ... + \beta_p \sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}.$$

Ahora, multiplicando por β',

$$\beta' X' X \beta = \begin{pmatrix} \beta_1 & \beta_2 & . & . & . & \beta_p \end{pmatrix} \begin{pmatrix} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{i1} + .. + \beta_p \sum_{i=1}^{n} x_{ip} \\ \beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + ... + \beta_p \sum_{i=1}^{n} x_{i1} x_{ip} \\ ... \\ \beta_0 \sum_{i=1}^{n} x_{ip} + \beta_1 \sum_{i=1}^{n} x_{i1} x_{ip} + ... + \beta_p \sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}$$

$$= \left(\beta_0^2 n + \beta_0 \beta_1 \sum_{i=1}^{n} x_{i1} + ... + \beta_0 \beta_p \sum_{i=1}^{n} x_{ip} \right) +$$

$$+ \left(\beta_0 \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_1^2 \sum_{i=1}^{n} x_{i1}^2 ... + \beta_1 \beta_p \sum_{i=1}^{n} x_{i1} x_{ip} \right) + ...$$

$$+ \left(\beta_0 \beta_p \sum_{i=1}^{n} x_{ip} + \beta_1 \beta_p \sum_{i=1}^{n} x_{i1} x_{ip} ... + \beta_p^2 \sum_{i=1}^{n} x_{ip}^2 \right).$$

Reordenando, queda:

$$\beta' X' X \beta = \left(\beta_0^2 n + 2\beta_0 \beta_1 \sum_{i=1}^{n} x_{i1} + ... + 2\beta_0 \beta_p \sum_{i=1}^{n} x_{ip} \right) +$$

$$+ \left(2\beta_0 \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_1^2 \sum_{i=1}^{n} x_{i1}^2 ... + 2\beta_1 \beta_p \sum_{i=1}^{n} x_{i1} x_{ip} \right) + ...$$

$$+ \left(2\beta_0 \beta_p \sum_{i=1}^{n} x_{ip} + 2\beta_1 \beta_p \sum_{i=1}^{n} x_{i1} x_{ip} ... + \beta_p^2 \sum_{i=1}^{n} x_{ip}^2 \right).$$

De esta forma,

$$\frac{\partial}{\partial\beta}\left(\beta X'X\beta'\right) = \begin{pmatrix} \frac{\partial}{\partial\beta_0}\left(\beta_0^2 n + 2\beta_0\beta_1\sum_{i=1}^{n} x_{i1} + ... + 2\beta_0\beta_p\sum_{i=1}^{n} x_{ip}\right) \\ \frac{\partial}{\partial\beta_1}\left(2\beta_0\beta_1\sum_{i=1}^{n} x_{i1} + \beta_1^2\sum_{i=1}^{n} x_{i1}^2 ... + 2\beta_1\beta_p\sum_{i=1}^{n} x_{i1}x_{ip}\right) \\ \vdots \\ \frac{\partial}{\partial\beta_p}\left(2\beta_0\beta_p\sum_{i=1}^{n} x_{ip} + 2\beta_1\beta_p\sum_{i=1}^{n} x_{i1}x_{ip} ... + \beta_p^2\sum_{i=1}^{n} x_{ip}^2\right) \end{pmatrix}$$

$$= \begin{pmatrix} 2n\beta_0 + 2\beta_1\sum_{i=1}^{n} x_{i1} + ... + 2\beta_p\sum_{i=1}^{n} x_{ip} \\ 2\beta_0\sum_{i=1}^{n} x_{i1} + 2\beta_1\sum_{i=1}^{n} x_{i1}^2 ... + 2\beta_p\sum_{i=1}^{n} x_{i1}x_{ip} \\ \vdots \\ 2\beta_0\sum_{i=1}^{n} x_{ip} + 2\beta_1\sum_{i=1}^{n} x_{i1}x_{ip} ... + 2\beta_p\sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}$$

$$= 2\begin{pmatrix} n\beta_0 + \beta_1\sum_{i=1}^{n} x_{i1} + .. + \beta_p\sum_{i=1}^{n} x_{ip} \\ \beta_0\sum_{i=1}^{n} x_{i1} + \beta_1\sum_{i=1}^{n} x_{i1}^2 + ... + \beta_p\sum_{i=1}^{n} x_{i1}x_{ip} \\ \cdots \\ \beta_0\sum_{i=1}^{n} x_{ip} + \beta_1\sum_{i=1}^{n} x_{i1}x_{ip} + ... + \beta_p\sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}$$

$$= 2X'X\beta.$$

Por lo tanto, se tiene:

$$\frac{\partial}{\partial\beta}\left(Y'Y - 2\beta'X'Y + \beta X'X\beta\right) = -2X'Y + 2X'X\beta. \tag{3.2.6}$$

Finalmente, para minimizar la función de mínimos cuadrados, se debe cumplir que $\frac{\partial S(\beta)}{\partial \beta} = 0$.

Pero, por la expresión (3.2.6), se observa:

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta.$$

Entonces buscando la solución para β de:

$$-2X'Y + 2X'X\beta = 0. \tag{3.2.7}$$

Sumando $2X'Y$ de ambos lados de (3.2.7),

$$2X'X\beta = 2X'Y. \tag{3.2.8}$$

Multiplicando ambos lados de (3.2.8) por $\frac{1}{2}$, se obtiene:

$$X'X\beta = X'Y. \tag{3.2.9}$$

Las soluciones de estas $p+1$ ecuaciones, denominadas *ecuaciones normales*, serán los estimadores $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$.

Por lo que, para encontrar la solución a las ecuaciones normales, se multiplica a ambos lados de (3.2.9) por la inversa de $X'X$, obteniendo que,

$$(X'X)^{-1}(X'X)\beta = (X'X)^{-1}X'Y.$$

Además, por definición de matriz inversa se sabe que, si A^{-1} es la matriz inversa de A, entonces $AA^{-1} = A^{-1}A = I$. Por tanto queda que:

$\beta = (X'X)^{-1}X'Y$, conforme lo que se quería demostrar.

Observación 3.2.2

La matriz $(X'X)^{-1}$ existirá si ninguna columna de la matriz X es una combinación lineal de cualquier otra columna, es decir, si las variables explicativas son linealmente independientes.

Finalmente, el modelo ajustado de regresión queda como:

$$\boxed{\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y.} \tag{3.2.10}$$

2.3. Intervalo de confianza para los coeficientes β_j

Para la construcción del intervalo se supone que los errores aleatorios tienen una distribución normal multivariada con vector de media 0 (con toda las entradas cero) y matriz de varianza-covarianza $\sigma^2\mathbb{I}$, donde $\sigma^2 > 0$ e $\mathbb{I}$ es la matriz identidad. De ahora en adelante se denotará esta distribución normal multivariada por $N(0, \sigma^2\mathbb{I})$.

Observación 3.3.1

Note que, dados β y X, se tiene que $X\beta$ es determinista, pues serán variables conocidas.
Por lo que, $\mathbb{E}(X\beta) = X\beta$ y $\mathbb{V}ar(X\beta) = 0$.

Lema 3.3.1

Dadas las variables explicativas X y los coeficientes de regresión β, la variable Y tiene distribución normal multivariada con vector de media $X\beta$ y matriz de varianza-covarianza $\sigma^2\mathbb{I}$.

Demostración

En términos del modelo de regresión, se puede escribir:

- Dado que $\mathbb{E}(\epsilon) = 0$ entonces,

$\mathbb{E}(Y) = \mathbb{E}(X\beta + \epsilon) = \mathbb{E}(X\beta) + \mathbb{E}(\epsilon) = X\beta.$

- Dado que $\mathbb{V}ar(\epsilon) = \sigma^2\mathbb{I}$ entonces,

$\mathbb{V}ar(Y) = \mathbb{V}ar(X\beta + \epsilon) = \mathbb{V}ar(X\beta) + \mathbb{V}ar(\epsilon) = \sigma^2\mathbb{I}.$

Por lo que, equivalentemente las variables de respuesta distribuyen $N(X\beta, \sigma^2\mathbb{I})$.

Lema 3.3.2

Si $\hat{\beta}$ es el estimador de β dado por la Proposición (3.2.2) y sea c_{jj} el j-ésimo elemento de la diagonal de la matriz $\sigma^2(X'X)^{-1}$, entonces $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}}$ se distribuye $N(0, 1)$.

Demostración

Se define a la matriz M como $M = (X'X)^{-1}X'$, y por la Proposición (3.2.2), se tiene que $\hat{\beta} = (X'X)^{-1}X'Y$, de tal forma que $\hat{\beta} = MY$.

Por propiedades de la normal multivariada, se sabe que si W se distribuye $N_p(\mu, \sum)$ y sea A una matriz de dimensión $p \times p$ no singular, entonces AW se distribuye $N_p(A\mu, A\sum A')$.

Entonces, por el Lema (3.3.1) y tomando $\hat{\beta} = MY$ se tendría lo siguiente:

Por un lado, $\mathbb{E}(\hat{\beta}) = \mathbb{E}(MY) = M\mathbb{E}(Y) = MX\beta$. Por otro lado, $\mathbb{V}ar(\hat{\beta}) = \mathbb{V}ar(MY) = M\sigma^2M'$, pero la matriz $\sigma^2\mathbb{I}$ es simétrica, por lo cual $M\sigma^2M' = \sigma^2MM'$.

Adicionalmente note que, $MX\beta = \underbrace{(X'X)^{-1}X'X}_{\mathbb{I}}\beta = \beta$

y que, $\sigma^2MM' = \sigma^2(X'X)^{-1}X'\left[(X'X)^{-1}X'\right]' = \sigma^2\underbrace{(X'X)^{-1}X'X}_{\mathbb{I}}\left[(X'X)^{-1}\right]'$

Por lo tanto, $\hat{\beta}$ se distribuye normal con vector de media β y matriz de varianza-covarianza $\sum = \sigma^2 \left[(X'X)^{-1}\right]'$. De esta forma, el j-ésimo elemento de la diagonal de la matriz $\sum$ es la varianza del estimador $\hat{\beta}_j$, mientras que el (ij)-ésimo elemento, $i \neq j$ (que no está de la diagonal) es la covarianza del entre $\hat{\beta}_i$ y $\hat{\beta}_j$.

Ahora, sabiendo que si W una matriz de dimensión $p \times 1$ tal que se distribuye $N_p(\mu, \sum)$, entonces todas las submatrices de W tienen una distribución normal. Así, el elemento w_i se distribuye $N_p(\mu_i, \sum_{ii})$, donde $\sum_{ii}$ es el i-ésimo elemento diagonal de la matriz $\sum$.

Por lo que, si $\sigma^2 c_{jj}$ es el j-ésimo elemento, entonces el estimador $\hat{\beta}_j$ se distribuirá normal con media β_j y varianza $\sigma^2 c_{jj}$, $j = 1, 2, .., p$.

Finalmente, estandarizando se obtiene que $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}}$ se distribuye $N(0, 1)$.

Lema 3.3.3

Si $\hat{\sigma}$ es el estimador de σ, entonces $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}$ se distribuye $\chi^2_{(n-p-1)}$.

Demostración

Ver Knight[1999].

Proposición 3.3.1

El intervalo del $100(1-\alpha)\,\%$ de confianza para $\hat{\beta}_j$ es:

$$\left(\hat{\beta}_j \quad \pm \quad t_{\frac{\alpha}{2},(n-p-1)} \sqrt{s.e.(\hat{\beta}_j)}\right).$$

Demostración

Para encontrar el intervalo de confianza para $\hat{\beta}_j$ se utiliza lo postulado por Knight[1999], sean x, z variables aleatorias tales que: x se distribuye $N(\mu, \sigma^2)$, y z se distribuye Chi-cuadrada con n grados de libertad. Entonces $\frac{x/\sigma^2}{\sqrt{y/n}}$ se distribuye t de student con n grados de libertad.

En consecuencia, se tiene que:

$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} = \frac{(\hat{\beta}_j - \beta_j)\sqrt{\sigma^2}}{\sqrt{\sigma^2 \hat{\sigma}^2 c_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}},$$

se distribuye t de student con (n-p-1) grados de libertad.

Por lo tanto, para $j = 1, 2, .., p$, el intervalo del $100(1-\alpha)\,\%$ de confianza para $\hat{\beta}_j$ queda como:

$$\left(\hat{\beta}_j \quad \pm \quad t_{\frac{\alpha}{2},(n-p-1)} \sqrt{\hat{\sigma}^2 c_{jj}}\right), \quad \text{donde} \sqrt{\hat{\sigma}^2 c_{jj}} \text{ es el error estándar del coeficiente } \hat{\beta}_j \text{ y se denota por } s.e.(\hat{\beta}_j).$$

2.4. Pruebas de hipótesis

2.4.1. Prueba de significancia de la regresión.

Por medio de esta prueba se puede determinar si al menos alguna de las variables independientes explica una cantidad significativa de la variación en la variable de respuesta.

Para esta prueba de significancia de la regresión, se tienen las siguientes hipótesis nula (H_0) y alternativa (H_a):

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_a : \text{Al menos un } \beta_j \neq 0$$

Si se rechaza la hipótesis nula, entonces se dice que al menos una de las variables X_j, $j = 1, 2, ..., p$ contribuye significativamente al modelo.

Ahora, para encontrar el estadístico de prueba se define lo siguiente para $i = 1, 2, .., n$:

- y_i el i-ésimo valor observado de la variable de respuesta.
- $\overline{y_i}$ el i-ésimo valor promedio de la variable predictora.
- $\hat{y}_i$ el i-ésimo valor predecido por el modelo de regresión.

Considere las siguientes variables:

i) Suma de Cuadrados Totales.

$$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

Mide la variación total de la variable de respuesta.

ii) Suma de Cuadrados por la Regresión.

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

Mide la variación en los valores observados de Y considerando la relación lineal entre los regresores y la variable de respuesta.

iii) Suma de Cuadrados por Error.

$$SCE = \sum (y_i - \hat{y}_i)^2.$$

Representa la variabilidad observada, es decir, aquello que no es explicado por el modelo de regresión.

Lema 3.4.1

Sean:

$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}$, el vector de los valores observados de la variable de respuesta.

$\overline{Y} = \begin{pmatrix} \overline{y} \\ \overline{y} \\ . \\ . \\ \overline{y} \end{pmatrix}$, el vector de valores promedio de la variable dependiente.

$\hat{Y} = \begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ . \\ . \\ \hat{y_n} \end{pmatrix}$, el vector de los valores predecidos por el modelo.

Entonces, se cumple lo siguiente:

i) $SCT = Y'Y + \overline{Y}'\overline{Y} - 2Y'\overline{Y}$.

ii) $SCR = \hat{\beta}'X'Y + \overline{Y}'\overline{Y} - 2\hat{\beta}'X'\overline{Y}$.

iii) $SCE = Y'Y + \hat{\beta}X'X\hat{\beta} - 2\hat{\beta}'X'Y$

Demostración

i) Note que, $(Y - \overline{Y})'(Y - \overline{Y}) = \begin{pmatrix} (y_1 - \overline{y}) & (y_2 - \overline{y}) & . & . & . & (y_n - \overline{y}) \end{pmatrix} \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ . \\ . \\ y_n - \overline{y} \end{pmatrix}$

$$= (y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + ... + (y_n - \overline{y})^2$$
$$= \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Entonces, la suma de cuadrados totales es equivalente a:

$$SCT = (Y - \overline{Y})'(Y - \overline{Y}).$$

Ahora, por propiedades de matrices se obtiene que:

$$SCT = (Y - \overline{Y})'(Y - \overline{Y}) = (Y' - \overline{Y}')(Y - \overline{Y}) = Y'Y + \overline{Y}'\overline{Y} - \overline{Y}'Y - Y'\overline{Y}.$$

Pero, $\overline{Y}'Y = \begin{pmatrix} \overline{y} & \overline{y} & . & . & . & \overline{y} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \overline{y}(y_1 + y_2 + ... + y_n)$

Así, $\overline{Y}'Y$ es una matriz de dimensión 1×1, por lo que su transpuesta será la misma matriz de dimensión 1×1, es decir, $\overline{Y}'Y = (\overline{Y}'Y)'$.

Además, de que $(\overline{Y}'Y)' = Y'(\overline{Y}')' = Y'\overline{Y}$.

Por tanto, se cumple que

$$SCT = Y'Y + \overline{Y}'\overline{Y} - Y'\overline{Y} - Y'\overline{Y} = Y'Y + \overline{Y}'\overline{Y} - 2Y'\overline{Y}.$$

ii) Note que, $(\hat{Y} - \overline{Y})'(\hat{Y} - \overline{Y}) = \begin{pmatrix} (\hat{y_1} - \overline{y}) & (\hat{y_2} - \overline{y}) & . & . & . & (\hat{y_n} - \overline{y}) \end{pmatrix} \begin{pmatrix} \hat{y_1} - \overline{y} \\ \hat{y_2} - \overline{y} \\ . \\ . \\ \hat{y_n} - \overline{y} \end{pmatrix}$

$$= (\hat{y_1} - \overline{y})^2 + (\hat{y_2} - \overline{y})^2 + ... + (\hat{y_n} - \overline{y})^2$$
$$= \sum_{i=1}^{n} (\hat{y_i} - \overline{y})^2.$$

Entonces, la suma de cuadrados por la regresión es equivalente a:

$$SCR = (\hat{Y} - \overline{Y})'(\hat{Y} - \overline{Y}).$$

Por la expresión (3.2.10) se tiene $\hat{Y} = X\hat{\beta}$, obteniendo así:

$$SCR = (X\hat{\beta} - \overline{Y})'(X\hat{\beta} - \overline{Y}).$$

Ahora, por propiedades de matrices se tiene que:

$$SCR = (X\hat{\beta} - \overline{Y})'(X\hat{\beta} - \overline{Y}) = ((X\hat{\beta})' - \overline{Y}')(X\hat{\beta} - \overline{Y}) = (\hat{\beta}'X' - \overline{Y}')(X\hat{\beta} - \overline{Y})$$
$$= \hat{\beta}'X'X\hat{\beta} - \hat{\beta}'X'\overline{Y} - \overline{Y}'X\hat{\beta} + \overline{Y}'\overline{Y}$$

Pero,

$$\overline{Y}'X\hat{\beta} = \begin{pmatrix} \overline{y} & \overline{y} & . & . & \overline{y} \end{pmatrix} \begin{pmatrix} \beta_0 + \beta_1 x_{11} + ... + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + ... + \beta_p x_{2p} \\ . \\ . \\ \beta_0 + \beta_1 x_{n1} + ... + \beta_p x_{np} \end{pmatrix} = \overline{y} \sum_{i=1}^{n} (\beta_0 + ... + \beta_p x_{ip})$$

Así, $\overline{Y}'X\hat{\beta}$ es una matriz de dimensión 1×1, por lo que su transpuesta será la misma matriz de dimensión 1×1, es decir, $\overline{Y}'X\hat{\beta} = (\overline{Y}'X\hat{\beta})'$.

Además, viendo que $(\overline{Y}'X\hat{\beta})' = (X\hat{\beta})'(\overline{Y}')' = \hat{\beta}'X'\overline{Y}$.

Por tanto, se cumple que

$$SCR = \hat{\beta}'X'Y + \overline{Y}'\overline{Y} - \hat{\beta}'X'\overline{Y} - \hat{\beta}'X'\overline{Y} = \hat{\beta}'X'Y + \overline{Y}'\overline{Y} - 2\hat{\beta}'X'\overline{Y}.$$

iii) Por el Lema (3.2.1) y la Observación (3.2.1) se tiene que:

$$SCE = \sum (y_i - \hat{y}_i)^2 = \epsilon'\epsilon = Y'Y - 2\beta'X'Y + \beta X'X\beta.$$

Proposición 3.4.1

Lo siguiente es válido:

$$SCR+SCE = SCT. \tag{3.4.1}$$

Demostración

Dado el Lema (3.4.1), se tiene que:

$SCT - SCR - SCE =$

$= (Y'Y + \overline{Y}'\overline{Y} - 2Y'\overline{Y}) - (\hat{\beta}'X'Y + \overline{Y}'\overline{Y} - 2\hat{\beta}'X'\overline{Y}) - (Y'Y + \hat{\beta}X'X\hat{\beta} - 2\hat{\beta}'X'Y)$

$= Y'Y + \overline{Y}'\overline{Y} - 2Y'\overline{Y} - \hat{\beta}'X'Y - \overline{Y}'\overline{Y} + 2\hat{\beta}'X'\overline{Y} - Y'Y - \hat{\beta}X'X\hat{\beta} + 2\hat{\beta}'X'Y$

$= \beta'X'Y - \beta'X'X\hat{\beta} - 2Y'\overline{Y} + 2\hat{\beta}'X'\overline{Y}$

$= (\beta'X'Y - \beta'X'X\hat{\beta}) + 2(\hat{\beta}'X'\overline{Y} - Y'\overline{Y}).$

Ahora, dado que el modelo ajustado está dado por $Y = X\hat{\beta}$, entonces por un lado se ve que:

$$\beta'X'Y - \beta'X'X\hat{\beta} = \beta'X'Y - \beta'X'Y = 0.$$

Por otro lado,

$$2(\hat{\beta}'X'\overline{Y} - Y'\overline{Y}) = 2((X\hat{\beta})'\overline{Y} - Y'\overline{Y}) = 2(Y'\overline{Y} - Y'\overline{Y}) = 0.$$

Por lo tanto, $SCT - SCR - SCE = 0$.

Es decir, $SCT = SCR + SCE$, como se quería demostrar.

La Proposición 3.4.1, ayudará a obtener el estadístico de prueba F, dado por:

$$F = \frac{SCR/p}{SCE/(n-p-1)}, \text{ el cual se distribuye } F_{(p,n-p-1)}.$$

Cabe destacar que el proceso completo para la obtención de este estadístico de prueba puede consultarse en el apéndice C.3, Montgomery et al. [2006].

Finalmente, para resumir mejor esta prueba se utilizará la siguiente tabla de Análisis de Varianza (ANOVA):

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	Estadístico de Prueba
Regresión	p	SCR	SCR/p	$F = \frac{SCR/p}{SCE/(n-p-1)}$
Error	n-p-1	SCE	SCE/(n-p-1)	
TOTAL	n	SCT		

Capítulo 3

Modelo de regresión lineal múltiple para el análisis de mecanismos genéticos

El planteamiento de determinados modelos estadísticos es de gran utilidad en el estudio de los mecanismos genéticos, es decir, en el análisis de la forma en que la transmisión de ciertos genotipos influye en rasgos observables específicos, también denomidados fenotipos.

En el presente capítulo se verá en particular la formulación de modelos de regresión múltiple para el estudio de la transimisión de características genéticas en un pedigree dado.

3.1. Notación

Continuando con lo planteado en el capítulo 2, la notación estará basada por una parte en lo propuesto por Elson y Stewart [1971], así como en los conceptos planteados por Bonney [1984].

Primero, por conveniencia y sin ninguna pérdida de generalidad se asumirá que dada una elección aleatoria de la población, los datos serán de un único pedigree.

Este pedigree estará conformado por un conjunto de N individuos con ancestros en el pedigree denotados por $X = (X_1, X_2, ..., X_N)$ y M cónyuges denotados por $Y = (Y_1, Y_2, ..., Y_M)$, posiblemente con $N \neq M$, ya que es probable que no todos los miembros estén emparejados o que algunos tengan más de una pareja. Por lo que, en conjunto los miembros y sus cónyuges serán agrupados en un vector de tamaño $n = N + M$ que se denota como $W = (W_1, W_2, ..., W_n)$.

El fenotipo del individuo W_i estará denotado por w_i. El vector de genotipos $\mathbf{w} = (w_1, w_2, ..., w_n)$, los k genotipos que causan variaciones en este fenotipo *(MG por sus siglas del inglés Major Gene)* estarán agrupados en el vector $\mathbf{u} = (u_1, u_2, ..., u_n)$.

Observación 4.1.1

Haciendo correspondencia con la notación utilizada en el capítulo 2, se tiene que W_i, coincide respectivamente a X_i para individuos con ancestros en el pedigree y a Y_i para los cónyuges sin ancestros dentro del pedigree.

Se tomará en cuenta la media poblacional de los fenotipos de los individuos que tuvieron genotipo u y ésta se denotará por μ_u, obteniendo así el vector $\mu_{\mathbf{u}} = (\mu_{u_1}, \mu_{u_2}, ..., \mu_{u_n})$.

Finalmente, se asumirá que el residual de los genotipos u es continuo, y se define como:

$$r = w - \mu_u. \tag{4.1.1}$$

Estos conceptos serán de gran utilidad para expresar la función de densidad del vector de fenotipos w en términos que realmente signifiquen algo biológicamente hablando. De esta forma, será necesario encontrar la probabilidad condicional de w dado u, es decir $\mathbb{P}(w \mid u)$ y también la probabilidad del vector de genotipos, $\mathbb{P}(u)$.

3.2. Probabilidad conjunta de los genotipos del pedigree

Para la construcción de los modelos estadísticos se considerarán las hipótesis propuestas por Bonney [1984], que son las dadas en el capítulo 2 y que también fueron consideradas por Elston y Stewart[1971] y Cannings et. al. [1978].

Tomando en cuenta las siguientes probabilidades:

•$\mathbb{P}(u)$ la probabilidad de que un individuo w, seleccionado aleatoriamente de la población, presente el genotipo u. Ésta corresponde a Ψ_u definida en el capítulo 2.

•$\mathbb{P}(u \mid u_F, u_M)$ la probabilidad condicional de que un individuo miembro del pedigree tenga genotipo u, dado que sus progenitores tuvieron genotipos u_F y u_M. Así se tiene que para el i-ésimo individuo cuyos progenitores están en el pedigre $\mathbb{P}(u_i \mid u_1, .., u_{i-1}) = \mathbb{P}(u_i \mid u_F, u_M)$, que es la hipótesis iv) dada en la sección 2.1.

•$\mathbb{P}(u \mid u_S)$ la probabilidad condicional de que un cónyuge tenga genotipo u, dado que su cónyuge tiene genotipo u_S. Pero, por el capítulo dos, se sabe que los genoptipos de los cónyuges al no estar emparentados, no dependen de otros genotipos, teniendo así que $\mathbb{P}(u_i \mid u_1, u_2, .., u_{i-1}) = \mathbb{P}(u_i) = \Psi_{u_i}$, para cónyuges sin ancestros en el pedigree.

Por lo que, la probabilidad conjunta de los genotipos del pedigree dado, podrá ser escrita como:

$$\mathbb{P}(u) = \mathbb{P}(u_1, u_2, ..., u_n) = \mathbb{P}(u_1)\mathbb{P}(u_2 \mid u_1)\mathbb{P}(u_3 \mid u_1, u_2)....\mathbb{P}(u_n \mid u_1, u_2, ..., u_{n-1}).$$

Pero, si además se define lo siguiente,

$$p_i = \begin{cases} \Psi_{u_i} = \mathbb{P}(u_i), \text{ si los ancestros del } i\text{-ésimo individuo no están en el pedigree,} \\ \mathbb{P}(u_i \mid u_F, u_M), \text{si los ancestros del } i\text{-ésimo individuo están en el pedigree,} \end{cases}$$

entonces se puede expresar la probabilidad conjunta de los genotipos como:

$$\mathbb{P}(u) = \prod_{i=1}^{n} p_i. \tag{4.2.1}$$

3.3. Distribución condicional de los residuales

Para obtener la distribución de los residuales, se asume que la probabilidad condicional del fenotipo w, dado el genotipo u, es decir, $\mathbb{P}(w \mid u)$ se distribuye normal con media μ_u y varianza $V = (\sigma_i \sigma_j \rho_{ij})$, con $i, j = 1, 2, ..., n$.

Además, se define lo siguiente para el i-ésimo individuo:

- σ_i^2 es la varianza condicional del fenotipo w_i, dado el genotipo u_i.
- μ_{u_i} es la media poblacional de fenotipos de los individuos con el genotipo u_i.
- ρ_{ij} la correlación entre los fenotipos w_i y w_j, del i-ésimo y j-ésimo individuo respectivamente. Esta correlación puede depender de la relación de parentezco entre individuos.

Lema 4.3.1

La distribución condicional del vector de residuales r, dado el vector de genotipos u, puede ser escrita como:

$$\mathbb{P}(r \mid u) = \prod_{i=1}^{n} \phi(z_i, v_i), \tag{4.3.1}$$

donde $\phi(z_i, v_i)$ es la distribución normal evaluada en z_i con media cero y varianza v_i.

Demostración

De la definición de residuales dada por la expresión (4.1.1), se puede observar que para obtener r únicamente se le está restando a w la media μ_u. Entonces, dado que la probabilidad $\mathbb{P}(w \mid u)$ es una normal con media μ_u y varianza $V = (\sigma_i \sigma_j \rho_{ij})$, se tiene $\mathbb{P}(r \mid u)$ se distribuye normal con media $0 = (\mu_u - \mu_u)$ y varianza V.

Ahora, con la finalidad de obtener una expresión más sencilla para $\mathbb{P}(r \mid u)$, se intentará escribirla como una multiplicación de densidades univariadas, para lo cual se van a transformar a los residuales en variables no relacionadas entre sí.

Indique por z_i los residuales ajustados, es decir, los resiudales transformados de forma que sean independientes y denote por β_i los coeficientes usuales de regresión, entonces se reescriben a los residuales de la siguiente forma:

$$z_i = \begin{cases} r_i & \text{para } i = 1. \\ r_i - \sum_{j=1}^{i-1} \beta_{ij} r_j & \text{para } i = 2, 3, ..., n. \end{cases}$$

Una propiedad especial de la distribución normal multivariada es que para aquellas variables cuya distribución conjunta es normal, la independencia y la correlación cero son condiciones equivalentes (Blitzstein, Hwang, 2014, p.313). Entonces, dado que $Cov(r_i, r_j) = 0$, pues los u_i son independientes, observe lo siguiente:

- Caso 1: $i \neq j, i = 1$

$$Cov(z_i, z_j) = Cov\left(r_1, r_j - \sum_{k=1}^{j-1} \beta_{jk} r_k\right) = Cov(r_1, r_j) - \sum_{k=1}^{j-1} \beta_{jk} Cov(r_1, r_k) = 0.$$

- Caso 2: $i \neq j, i \neq 1$

$$\begin{aligned} Cov(z_i, z_j) &= Cov\left(r_i - \sum_{k=1}^{i-1} \beta_{ik} r_k, r_j - \sum_{m=1}^{j-1} \beta_{jk} r_m\right) \\ &= Cov(r_i, r_j) - \sum_{m=1}^{j-1} \beta_{jm} Cov(r_i, r_m) - \sum_{k=1}^{i-1} \beta_{ik} Cov(r_k, r_j) - \sum_{k=1}^{i-1} \sum_{m=1}^{j-1} \beta_{jm} Cov(r_k, r_m) = 0. \end{aligned}$$

Además, dado que $Cov(z_i, z_j) = 0$, entonces $\rho_{ij} = 1$. Por lo que el i-ésimo residual ajustado tendrá varianza σ_i^2, que es el i-ésimo elemento de la matriz V y que estará denotado por v_i.

Por lo tanto, bajo la suposición de que los residuales se distribuyen normal con media cero y varianza V, los residuales ajustados son independientes y se distribuyen normal con media cero y varianza v_i.

Finalmente, se tendría que:

$$\mathbb{P}(r \mid u) = \prod_{i=1}^{n} \mathbb{P}(z_i \mid u) = \prod_{i=1}^{n} \phi(z_i, v_i), \tag{4.3.1}$$

donde $\phi(z_i, v_i)$ es la distribución normal evaluada en z_i con media cero y varianza v_i.

3.4. Covarianzas

De acuerdo con Bonney [1984], es necesario considerar las covarianzas de los cónyuges con los demás miembros del pedigre, las covarianzas que existen dentro de una hermandad y las que existen entre hermanos. Esto es debido a que los residuales únicamente representan los efectos de la información genética transmitida y de los factores ambientales, tanto aleatorios como no aleatorios.

3.4.1. Covarianza conyugal.

Como se vio en el capítulo 2, generalmente se asume que para el i-ésimo individuo, con sus ancestros en el pedigree, su cónyuge tiene su residual r_i no correlacionado con los residuales de los demás miembros de generaciones anteriores del pedigree, excepto probablemente con el residual r_s de su respectivo cónyuge.

Entonces, se tendría que:

$$z_i = r_i - \rho_{FM}\frac{\sigma_i}{\sigma_S} r_s = (w_i - \mu_{u_i}) - \rho_{FM}\frac{\sigma_i}{\sigma_S}(w_S - \mu_{u_S}), \tag{4.4.1}$$

donde es la desviación estándar del residual r_S y w_S es el fenotipo del cónyuge del i-ésimo individuo. Mientras que ρ_{FM} es la correlación residual entre el padre y la madre, en este caso, la conyugal.

3.4.2. Covarianza entre miembros del pedigree.

Por la estructura ya construida a lo largo de los capítulos pasados, se sabe que el residual del i-ésimo individuo únicamente dependerá de los residuales de sus progenitores, los cuales a su vez dependerán de los residuales de los abuelos del i-ésimo individuo. De esta forma, sea ρ_{PO} la correlación de los residuales entre padres y descendencia, entonces la correlación de los residuales entre abuelos y descendencia estará dada por ρ^2_{PO}.

3.4.3. Covarianza dentro de un conjunto de hermanos.

Existen diversas causas de la covarianza entre hermanos, por esta razón a continuación se describirán tres diferentes modelos para un pedigree dado.

Observación 4.4.1

Antes de exponer los modelos, se ve el procedimiento realizado por Bonney[1984]. En primer lugar, para el padre, la madre y H hijos, denotados por $W_F, W_M, W_1, W_2, ..., W_H$ respectivamente, se define lo siguiente:

- σ^2_F a la varianza condicional del padre, denotado por W_F.
- σ^2_M a la varianza condicional de la madre, que se denota por W_M.
- σ^2_{Fi} a la desviación estándar entre el padre y el i-ésimo hijo, con $i = 1, 2, ..., H$.
- σ^2_{Mi} a la desviación estándar entre la madre y el i-ésimo hijo, con $i = 1, 2, ..., H$.
- σ^2_i a la varianza condicional de W_i, con $i = 1, 2, ..., H$.
- ρ_{FM} a la correlación entre padre y madre.
- ρ_{FH} a la correlación entre el padre y su descendencia, teniendo en particular para el i-ésimo de sus hijos ρ_{Fi}.
- ρ_{MH} a la correlación entre la madre y su descendencia, teniendo en particular para el i-ésimo de sus hijos ρ_{Mi}.

De tal forma que la matriz de varianzas de los residuales ajustados para $W_F, W_M, W_1, W_2, ..., W_H$ puede ser escrita como:

$$V = \left[\begin{array}{cc|ccccc} \sigma^2_F & \sigma_F\sigma_M\rho_{FM} & \sigma_F\sigma_1\sigma_{F1} & \sigma_F\sigma_2\sigma_{F2} & ... & \sigma_F\sigma_H\sigma_{FH} \\ \sigma_F\sigma_M\rho_{FM} & \sigma^2_M & \sigma_M\sigma_1\sigma_{M1} & \sigma_M\sigma_2\sigma_{M2} & ... & \sigma_M\sigma_H\sigma_{MH} \\ \hline \sigma_F\sigma_1\sigma_{F1} & \sigma_M\sigma_1\sigma_{M1} & \sigma^2_1 & \sigma_1\sigma_2\rho_{12} & ... & \sigma_1\sigma_H\rho_{1H} \\ . & . & . & & & \\ . & . & . & & & \\ \sigma_F\sigma_H\sigma_{FH} & \sigma_M\sigma_H\sigma_{MH} & \sigma_1\sigma_H\rho_{1H} & \sigma_2\sigma_H\rho_{2H} & ... & \sigma^2_H \end{array}\right]$$

Por lo que, $V = \left[\begin{array}{c|c} V_F & V_{PH} \\ \hline V_{HP} & V_H \end{array}\right]$.

Entonces, se tiene que:

- La matriz de coeficientes parciales de la descendencia (H hijos), será $V_{HP}V_P^{-1}$. Por lo que, la i-ésima columna será $(\beta_{iF}\beta_{iM})$, donde:

$$i)\beta_{iF} = \frac{\rho_{Fi} - \rho_{FM}\rho_{Mi}}{1 - \rho_{FM}^2}\left(\frac{\sigma_i}{\sigma_F}\right). \tag{4.4.2}$$

$$ii)\beta_{iM} = \frac{\rho_{Mi} - \rho_{FM}\rho_{Fi}}{1 - \rho_{FM}^2}\left(\frac{\sigma_i}{\sigma_M}\right).$$

- La matriz de varianzas de los residuales ajustados de los hijos, dados los residuales ajustados de los padres está dada por $V_{HP}V_P^{-1}V_{PH}$. En consecuencia, el elemento de la i-ésima fila y la j-ésima columna, con $i, j = 1, 2, ..., n$, será $\sigma_i\sigma_j\delta_{ij}$, donde:

$$\delta_{ij} = \frac{(\rho_{Fi} - \rho_{FM}\rho_{Mi})\rho_{Fj} + (\rho_{Mi} - \rho_{FM}\rho_{Fi})\rho_{Mj}}{1 - \rho_{FM}^2} \tag{4.4.3}$$

De esta forma, se tiene que la covarianza condicional de w_i y w_j, dados los residuales de los padres, r_F y r_M, será $\sigma_i\sigma_j\rho_{ij} - \sigma_i\sigma_j\delta_{ij}$. Entonces la correlación parcial entre w_i y w_j, dados r_F y r_M, está dada por:

$$\rho_{\cdot ij} = \frac{\rho_{ij} - \delta ij}{\sqrt{1 - \delta_{ii}}\sqrt{1 - \delta_{jj}}}.$$

A partir de esta última expresión se puede obtener la correlación entre los residuales ajustados del i-ésimo y el j-ésimo hijo, para ello primero se multiplicarán ambos lados por $\sqrt{1 - \delta_{ii}}\sqrt{1 - \delta_{jj}}$,

$$\rho_{ij} - \delta ij = \rho_{\cdot ij}\sqrt{1 - \delta_{ii}}\sqrt{1 - \delta_{jj}}.$$

Posteriormente, sumando δij de ambos lados de la última expresión, para así obtener que:

$$\rho_{ij} = \rho_{\cdot ij}\sqrt{1 - \delta_{ii}}\sqrt{1 - \delta_{jj}} + \delta ij.$$

Cabe destacar que δ_{ii} es la proporción de la varianza de los residuales ajustados del i-ésimo hijo explicada por la regresión sobre los residuales de los padres.

Modelos de tipo A.

Estos modelos son factibles cuando las causas de las correlaciones entre hermanos son principalmente: transmisión biológica, transmisión cultural y los ambientes parentales compartidos. Por lo que, se asumirá que las correlaciones residuales entre hermanos se deben únicamente a que tienen los mismos padres.

Para el i-ésimo hermano, miembro del pedigre, definiendo lo siguiente:

- w_{iF}, w_{iM} los fenotipos de su padre y su madre respectivamente.
- μ_{iF}, μ_{iM} la media poblacional de los fenotipos que tuvieron genotipo u, para su padre y su madre respectivamente.
- r_{iF}, r_{iM} los residuales de su padre y su madre respectivamente.

Entonces, el residual ajustado de este individuo i, únicamente estará ajustado por los residuales de su padre y su madre, obteniendo así:

$$z_i = r_i - \beta_{iF} r_F - \beta_{iM} r_M = (w_i - \mu_{u_i}) - \beta_{iF}(w_F - \mu_F) - \beta_{iM}(w_M - \mu_M) \tag{4.4.4}$$

Como se vio en la sección 3.3, se puede interpretar a los coeficientes parciales de este modelo como: β_{iF} es el cambio en z_i, asociado a un cambio de una unidad en r_F, manteniendo r_M constante, e igualmente β_{iM} es el cambio en z_i, asociado a un cambio de una unidad en r_M, manteniendo r_F constante.

Ahora, viendo el caso en que todas las varianzas σ^2 son iguales y las correlaciones entre padres y descendencia son iguales, es decir, el caso particular de la Observación (4.4.1), donde:
$\rho_{PH} = \rho_{MH} = \rho_{FH}$.

Así, sustuyendo ρ_{PH} en (4.4.2), se obtiene que:

i) $\beta_{iF} = \beta_{iM} = \frac{\rho_{PH} - \rho_{FM}\rho_{PH}}{1-\rho_{FM}^2} = \frac{\rho_{PH}(1-\rho_{FM})}{(1+\rho_{FM})(1-\rho_{FM})} = \frac{\rho_{PH}}{1+\rho_{FM}}$.

ii) Además sustituyendo ρ_{PO} en (4.4.3), se tiene que la varianza de z_i está dada por: $v_i = \sigma^2(1-\delta)$, donde:

$\delta = \frac{(\rho_{PH} - \rho_{FM}\rho_{PH})\rho_{PH} + (\rho_{PH} - \rho_{FM}\rho_{PH})\rho_{PH}}{1-\rho_{FM}^2} = \frac{2(\rho_{PH} - \rho_{FM}\rho_{PH})\rho_{PH}}{(1+\rho_{FM})(1-\rho_{FM})} = \frac{2\rho_{PH}^2(1-\rho_{FM})}{(1+\rho_{FM})(1-\rho_{FM})} = \frac{2\rho_{PH}^2}{(1+\rho_{FM})}$
es la proporción de la varianza de los residuales explicada por la regresión basada únicamente en los padres.

Asimismo, como v_i y σ^2 tienen que ser positivas y como $v_i = \sigma^2(1-\delta)$, entonces se debe cumplir que $(1-\delta) \geq 0$, es decir, $0 \leq \delta < 1$. De tal forma que quede que, $0 \leq \frac{2\rho_{PH}^2}{1+\rho_{FM}} < 1$.

Multiplicando esta última expresión por $\frac{1+\rho_{FM}}{2}$, se obtiene que $0 \leq \rho_{PH}^2 < \frac{1+\rho_{FM}}{2}$.

Por otro lado, si se denota como ρ_{SS} a la correlación entre los residuales r_i, r_j de los hermanos i, j, y se define como $\rho*_{SS}$ a la correlación condicional entre los residuales ajustados z_i, z_j de los hermanos i, j dados los residuales de los padres, se obtiene que:

$$\rho_{SS} = \delta + (1-\delta)\rho*_{SS} \tag{4.4.5}$$

De la expresión (4.4.5), se puede ver que si $\rho*_{SS} = 0$, entonces $\rho_{SS} = \delta$.

Así que, si $\rho*_{SS}$ es diferente de cero, se puede decir que δ es la porción de la correlación de los residuales de los hermanos debida a que tienen los mismos padres. Mientras que $(1-\delta)\rho*_{SS}$ es la porción de la correlación que no se debe a que tengan los mismos padres.

De la misma forma que para los modelos de clase A, las correlaciones entre hermanos únicamente se deben a que comparten progenitores, entonces se tendrá la restricción $\rho_{SS} = \delta$. Por lo que para este caso del modelo A, se tendrían que tomar en cuenta las siguientes consideraciones:

1) Las correlaciones entre hermanos son iguales pero no pueden ser negativas.

2) Si el padre y la madre no están relacionados ($\rho_{FM} = 0$), entonces la correlación entre entre hermanos varía $2\rho_{PH}^2$.

3) Al incrementar ρ_{FM}, la correlación entre hermanos decrece, pero no puede ser menor a ρ_{PH}^2.

4) Una correlación entre progenitores que es grande, pero negativa puede llevar a una alta correlación entre hermanos.

Por otro lado, se dice que un " modelo equivariante" es áquel en el se asume que todas las varianzas residuales son iguales dados los genotipos. Este tipo de modelos reduce considerablemente el número de parámetros en el modelo y toma en cuenta la constancia de la variabilidad genética, Mather [1942].

Modelos de tipo B.

En estos modelos, las correlaciones residuales entre hermanos no sólo se deben a los progenitores que comparten, sino que toman en cuenta a sus hermanos mayores.

En particular, las correlaciones residuales entre una generación de hermanos estarán determinadas de tal forma que, dados los residuales de los progenitores y los de los primeros m hermanos mayores, los residuales de los demás hermanos serán independientes.

A continuación, se verá que ocurre cuando las dependencias entre los residuales de una generación de hermanos se debe a que comparten padre, madre y un hermano mayor, es decir, el caso en el que $m = 1$.

En este caso, el residual ajustado z_1 para el primer hermano es el mismo que el obtenido en el modelo A, mientras que para $i = 2, 3, .., n$ se tendría que:

$$\begin{aligned} z_i &= r_i - \beta_{iF} r_F - \beta_{iM} r_M - \beta_{i1} r_1 \\ &= (w_i - \mu_{u_i}) - \beta_{iF}(w_F - \mu_F) - \beta_{iM}(w_M - \mu_M) - \beta_{i1}(w_1 - \mu_1). \end{aligned} \tag{4.4.6}$$

Y aunque los residuales de los hermanos están ajustados por los padres, en este tipo de modelos las variaciones ajustadas aún están correlacionadas.

En consecuencia, se tiene el caso particular de la Observación 4.4.1, en el que $\rho_{FO} = \rho_{Fi}$ y $\rho_{MO} = \rho_{Mi}$, por lo que si se sustituye esto último en 4.4.2, se obtendría que:

i) $\alpha_F = \frac{\rho_{FO} - \rho_{FM}\rho_{MO}}{1-\rho_{FM}^2}$ y $\alpha_M = \frac{\rho_{MO} - \rho_{FM}\rho_{FO}}{1-\rho_{FM}^2}$.

Ahora, sustituyendo ρ_{FO} y ρ_{MO} en 4.4.3, se tiene que:

ii) $\delta = \frac{(\rho_{FO}-\rho_{FM}\rho_{MO})\rho_{FO}+(\rho_{MO}-\rho_{FM}\rho_{FO})\rho_{MO}}{1-\rho_{FM}^2}$
$= \frac{(\rho_{FO}^2-\rho_{FM}\rho_{MO}\rho_{FO})+(\rho_{MO}^2-\rho_{FM}\rho_{MO}\rho_{FO})}{1-\rho_{FM}^2} = \frac{\rho_{FO}^2+\rho_{MO}^2-2\rho_{FM}\rho_{MO}\rho_{FO}}{1-\rho_{FM}^2}$

Por lo que, para los hermanos restantes, $1 = 2, 3, .., n$ el residual ajustado estará dado por:

$$\begin{aligned} z_i &= r_i - \alpha_F r_F - \alpha_M r_M - \tfrac{\rho_{i1}-\delta}{1-\delta}(y_1 - \alpha_F r_F - \alpha_M r_M) \\ &= r_i - \alpha_F r_F(1 - \tfrac{\rho_{i1}-\delta}{1-\delta}) - \alpha_M r_M(1 - \tfrac{\rho_{i1}-\delta}{1-\delta}) + r_1(1 - \tfrac{\rho_{i1}-\delta}{1-\delta}) \\ &= r_i - \beta_{iF} r_F - \beta_{iM} r_M - \beta_{i1} r_1, \end{aligned}$$

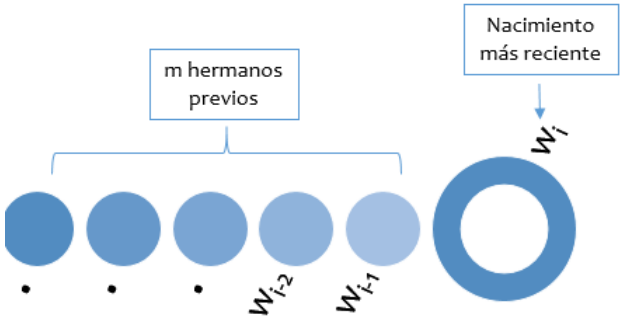
donde:

- $\beta_{iF} = \alpha_F(1 - \frac{\rho_{i1}-\delta}{1-\delta}) = \alpha_F(\frac{1-\delta-\rho_{i1}+\delta}{1-\delta}) = \alpha_F(\frac{1-\rho_{i1}}{1-\delta})$.
- $\beta_{iM} = \alpha_M(1 - \frac{\rho_{i1}-\delta}{1-\delta}) = \alpha_M(\frac{1-\delta-\rho_{i1}+\delta}{1-\delta}) = \alpha_M(\frac{1-\rho_{i1}}{1-\delta})$.
- $\beta_{i1} = \frac{\rho_{i1}-\delta}{1-\delta}$.

Modelos de tipo C.

Esta clase de modelos se basa en la idea intuitiva para los hermanos de que cuanto más cercano el orden de sus nacimientos, más relacionados están. Entonces, las correlaciones residuales entre una generación de hermanos, condicionadas por los residuales de los progenitores, dependerán de los m hermanos previos y porteriores conforme al orden de su respectivo nacimiento.

Utilizando una línea del tiempo tendríamos algo como:



Entonces, para los residuales ajustados del i-ésimo hermano, sólo será necesario ajustar los residuales de los m hermanos que los preceden.

Ahora, si se considera el caso en el que después de ajustar los residuales debido a los progenitores, sólo un hermano precede en orden de nacimiento, es decir $m = 1$, y denotando como ρ a la correlación residual entre w_i y w_{i-1}, la correlación residual entre w_i y w_{i-2}, es decir, entre dos hermanos separados por el nacimiento del hermano w_{i-1}, se escribirá como ρ^2.

De esta forma, la correlación residual, después del ajuste debido a los progenitores, para dos hermanos separados por t nacimientos estará dada por ρ^{t+1}.

Finalmente, para los residuales ajustados se tendrá lo mismo que en los modelos de tipo B para z_1, mientras que para $i = 2, 3, .., n$ obteniendo:

$$\begin{aligned} z_i &= r_i - \beta_{iF} r_F - \beta_{iM} r_M - \beta_{i,i-1} r_{i-1} \\ &= (w_i - \mu_{u_i}) - \beta_{iF}(w_F - \mu_F) - \beta_{iM}(w_M - \mu_M) - \beta_{i,i-1}(w_{i-1} - \mu_{i-1}). \end{aligned} \tag{4.4.7}$$

Estos coeficientes parciales de regresión son los mismos que en (4.4.6), cambiando β_{i1} por $\beta_{i,i-1}$.

Conclusiones

Los modelos estadísticos de regresión lineal múltiple presentados en esta tesina se basaron en el análisis de rasgos observables específicos, medidos en pedigrees que fueron elegidos aleatoriamente y a partir de los cuales se explica el mecanismo de ciertos genotipos.

Previo a esto, fue necesario establecer modelos de probabilidad que permitieron analizar los mecanismos genéticos de un pedigree dado, definiendo así su respectiva función de verosimilitud para que fuera posible especificar la transmisión genética de un individuo a su descendiente, y así elucidar como la transmisión de ciertos genotipos influye en la expresión de rasgos observables particulares.

Aunque no se profundizó en las aplicaciones que estos modelos tienen en la generación de nuevas hipótesis científicas para el desarrollo de un enfoque de "medicina personalizada", debe señalarse que estos modelos, que son computacionalmente factibles, son escenciales para comprender e interpretar los mecanismos genéticos.

Por lo cual, se podría decir que los modelos considerados forman parte de la base en la comprensión e interpretación de la gransmisión de la información heredada de generación en generación, y precisamente este entendimiento de la transmisión genética trae consigo diversas aplicaciones en el mejoramiento de la calidad de vida, pues pueden ayudar a la detección de variaciones genéticas que controlan rasgos específicos o enfermedades para que así sea posible elaborar nuevas estrategias en la prevención y el tratamiento de estos padecimientos.

Referencias

[1] Blitzstein, J. K., Hwang, J. (2014). *Introduction to probability.* Florida: Chapman & Hall/CRC.

[2] Bonney, G. E. (1984). *On the statistical determination of major gene mechanisms in continuous human traits: Regressive models.* Am. J. Med. Genet., 18(4): 731-749, DOI: 10.1002/ajmg.1320180420.

[3] Cannings, C., Thompson, E.A., Skolnick, M.H. (1978). *Probability functions on complex pedigrees.* Advances in Applied Probability, 10(1): 26–61. DOI: 10.1017/S0001867800029475.

[4] Elston, R. C., Stewart, J. (1971). *A general model for the genetic analysis of pedigree data.* Human Heredity, 21(6): 523-542, DOI: 10.1159/000152448.

[5] Knigth, K. (1999). *Mathematical statistics.* Florida: Chapman & Hall/CRC.

[6] Mather, K. (1942) *The balance of polygenic combinations.* J. Genet. 43: 309-336.

[7] Montgomery, D.C., Peck, E.A., Vining, G.G. (2006). *Introduction to linear regression analysis.* 4ta ed. New Jersey: Wiley.