UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Aplicación del modelo de regresión logística en mediciones de registros vocales para el diagnóstico de la enfermedad de Parkinson.

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I O

PRESENTA:
JONNATHAN GUTIÉRREZ ORTIZ

DIRECTORA DE TESIS:
DRA. LIZBETH NARANJO ALBARRÁN

Ciudad Universitaria, 2017.

Agradecimientos

A Dios mi Salvador, por todo. Por darme la vida y la guía para vivirla en Su palabra. Por la gracia y el perdón en Cristo. Por Su dulce amor bendito. Por guardar mi entrada y mi salida. Y por Su mano fiel que ha sostenido mi vida, en la dicha o el dolor. Por ayudarme a concluir una licenciatura y el presente trabajo de tesis para Su gloria. Lector: ¡busca a Dios con todo tu corazón! Él ha dicho: "Buscadme, y viviréis".

A mis padres, Rodolfo y Clara, por su amor y ayuda inconmensurables a través de los años. De manera muy especial a mi querida madre, por su apoyo incondicional hacia mi hermana y hacia mí, por habernos dado sustento siempre, incluso en los momentos en que era más difícil, para salir adelante y "llegar al menos a la luna", como suele decir. Le estaré eternamente agradecido. Este trabajo es fruto de su esfuerzo.

A mi hermana Jenny, con quien llegué al mundo, y con quien espero compartir muchos años más a su lado. Su tenacidad en el estudio muchas veces me ha redargüido y exhortado a mejorar. ¡Te quiero mucho Jenny!

A mis tíos, que me han sido por otros padres durante los últimos años de mi vida. A mi tío Óscar a quien he llegado a querer grandemente y sé que él también a mí. Le agradezco su enseñanza y constante ayuda, por lo cual le estoy en deuda. Este pequeño logro también es suyo. Y a mi tía Rosaura por aceptarme como a hijo.

A mi primo Gabriel, a quien quiero tanto como a un hermano en la carne. Espero sinceramente que nuestra amistad nunca cambie, sino sólo para mejorar. Y a mi pequeña prima Abigail, la niña de mis ojos, por alegrarme con su niñez.

Al resto de mi familia también le debo esto. En particular, lo dedico a mi abuelito Germán, a quien extraño y admiro, y que padeció la enfermedad de Parkinson.

A mis hermanas y hermanos en Cristo por sus oraciones, amistad y la fraternidad en que tanto gozo encuentro. Son muchos hermanos, gracias a Dios, pero agradezco en particular la amistad de Ubi, Noemí y Nury.

A Andrea, Fernando, Norma, Alex, Lupita, Yessi, Melissa, Aileén, Andrés, Erick, Carolina y Wendy, por su amistad y apoyo durante la licenciatura.

A mis (¡muchos!) maestros del CELE que tanto me enseñaron. En especial a mi querida maestra Laura Gasparyan Jachatrian, donde quiera que esté, a Victoria Nistor, y a mi muy amable maestra Anny Chryssomalakou que también me quiere mucho.

A mis entrañables profesores Javier Páez, Ana Meda, Margarita Chávez Cano y Guillermina Eslava, entre otros que despertaron en mí el gusto por las matemáticas y la estadística, para quienes guardo una profunda estima, a pesar de no ser, por mucho, uno de sus mejores alumnos.

A los sinodales, por las correcciones amablemente hechas a este trabajo.

Y finalmente y de manera muy personal, agradezco a mi directora de tesis, la Doctora Lizbeth Naranjo Albarrán por su ayuda constante y dirección a lo largo de este proceso. Disfruté mucho realizar este trabajo con su apoyo. Cada instrucción me fue de gran auxilio. Tomo prestadas las palabras que un buen amigo escribió: “lo que más le agradezco es haberme permitido ser su amigo, una afirmación arriesgada de mi parte, pero siempre me hizo sentir así.” Mi profundo agradecimiento, respeto y admiración estarán siempre presentes para ella.

Índice general

Capítulo 1

Resumen

La enfermedad de Parkinson es una patología que, en promedio, es diagnosticada entre 10 y 20 años después de su inicio, provocando así el avance de los indeseables efectos neurodegenerativos de la enfermedad sobre la persona para el momento en que ésta es diagnosticada. La complicación en el diagnóstico se encuentra relacionada con los síntomas de la enfermedad, los cuales son mayormente motores, y que, por lo tanto, pueden confundirse con los ocasionados por otros padecimientos o incluso la edad. El método de diagnóstico tradicional consiste en una serie de pruebas motoras en las cuales se reflejan la presencia y severidad de los síntomas de la enfermedad, evaluados con la *Escala Unificada de Clasificación de la Enfermedad de Parkinson* UPDRS, por sus siglas en inglés (*Unified Parkinson's Disease Rating Scale*), lo que resulta en una evaluación subjetiva de la capacidad del individuo para realizar ciertas tareas [15].

Uno de los primeros sistemas motores que la enfermedad ataca es el sistema del habla, de modo tal que el 90 % de las personas con la enfermedad padecen trastornos del habla [16], los cuales al ser medidos acústicamente poseen parámetros sensitivos que son detectables hasta cinco años antes de la edad *promedio* del diagnóstico de la enfermedad [7].

Por ello, se tomaron registros vocales de 40 personas con la enfermedad de Parkinson, pertenecientes a la Asociación Regional para la Enfermedad de Parkinson en Extremadura (España) [12], y de 40 personas dentro de un grupo control de individuos sin la enfermedad de Parkinson. A partir de tales registros (grabaciones) se obtuvieron en total 44 mediciones de distintas características acústicas para cada persona y estas características se estudiarán para conocer cuáles de ellas permiten diagnosticar correctamente la presencia de la enfermedad de Parkinson por medio del modelo de regresión logística.

En este primer capítulo se presenta una primera introducción a la motivación y objetivo de este trabajo de investigación (y se ofrece una explicación más extensa en el capítulo

quinto). También se describe de forma preambular el contenido de cada uno de los capítulos de este trabajo.

En el segundo capítulo se establecen algunas nociones preliminares utilizadas al trabajar con variables aleatorias que se encuentran asociadas a una respuesta, y la forma en que pueden estimarse probabilidades a partir de datos muestrales cuando se hacen separaciones categóricas como en el caso de las tablas de contingencia. Y a partir de esta última, se obtienen la sensibilidad y especificidad del método de diagnóstico en estudio, el riesgo relativo y la razón de momios de la obtención de cierto diagnóstico.

En el tercer capítulo se presentan teóricamente la descomposición espectral y el análisis de los componentes principales de una matriz. Así como la estimación de la descomposición espectral y el análisis de los componentes principales de una matriz de covarianzas muestral y el análisis de componentes principales comunes para casos en que existen distintos grupos o submuestras.

En el capítulo cuarto se presenta una introducción a los modelos lineales generalizados, en particular a los modelos de respuesta binaria y las ligas canónicas usualmente utilizadas en este tipo de modelos. También se explica la construcción del modelo de regresión logística y la obtención de los coeficientes de regresión, así como algunas pruebas de hipótesis para los parámetros y de bondad de ajuste del modelo, así como su interpretación.

En el quinto capítulo se ofrece una introducción contextual a modo de justificación del método propuesto para diagnosticar la enfermedad de Parkinson, se detalla la información acerca de los materiales y métodos usados, así como una explicación de las características de las variables en la base de datos. Se añade una explicación de la aplicación hecha con el método de componentes principales comunes, la cual se realizó con el software estadístico R. A partir de la selección de los componentes principales, la obtención del modelo de regresión logística, así como su interpretación. Se incluyen también las pruebas de hipótesis de los parámetros y las pruebas de bondad de ajuste del modelo. Se presentan la sensibilidad y especificidad del modelo obtenido, además de su interpretación, el riesgo relativo y la razón de momios al usar este método de diagnóstico. Una sección, a modo de resumen, de los resultados más importantes encontrados en este trabajo, así como un análisis cualitativo del método de diagnóstico utilizado por medio del análisis de la curva ROC. Y una última sección de validación cruzada, donde se reportan únicamente las medias de los resultados obtenidos en las cien iteraciones del proceso.

En el capítulo sexto, se añade un resumen de las posibles técnicas o métodos alternativos para la resolución de un problema como el aquí planteado, y se señalan algunas

debilidades de este trabajo, tal como el problema de abordar medidas repetidas y de reemplazar estas medidas repetidas por su media.

Finalmente, se presentan dos anexos con algunas propiedades de matrices y de funciones de distribución, así como algunas funciones liga y el desarrollo de la versión multivariada del método Delta, además del código de R utilizado en este trabajo.

Capítulo 2

Preliminares

2.1. Probabilidades condicionales

Existen diversos tratamientos matemáticos aplicados a fenómenos de la vida diaria de los cuales, luego de definir un evento específico, es posible encontrar su probabilidad de ocurrencia (e.g. obtener águila en el lanzamiento de una moneda). Sin embargo, en otras ocasiones nos encontramos más interesados en conocer los factores que permiten obtener ese resultado.
En la práctica, los propósitos de encontrar dichos factores son muy variados, pero en general, se precisa saber más que la ocurrencia o no de un evento, a saber, con qué probabilidad ocurrirá tomando en cuenta las condiciones dadas. En la teoría, la siguiente función ofrece una respuesta a nuestro cuestionamiento.
Sean X y Y variables aleatorias y sea $\mathfrak{D}(X)$ el dominio de X, entonces $\mathbb{P}(Y \mid X = x)$, para toda $x \in \mathfrak{D}(X)$, es la probabilidad condicional de la respuesta estudiada Y, dada la presencia de los valores de entrada de X.

2.2. Tablas de contingencia

Una *tabla de contingencia* es un arreglo matricial de dimensión $m \times n$, o tabla de contingencia de $m \times n$, donde el ij-ésimo elemento corresponde a la probabilidad de que X tome el valor i y Y tome el valor j. Sea π_{ij} el ij-ésimo elemento de una tabla de contingencia de estructura probabilística, entonces la distribución conjunta es $\mathbb{P}(X = i, Y = j) = \pi_{ij}$. La distribución marginal de dos vectores cualesquiera, renglón y columna respectivamente, se denota como $\pi_{i\bullet}$ y $\pi_{\bullet j}$, para $i \in \{1, \ldots, m\}$ y $j \in \{1, \ldots, n\}$, donde

$$\pi_{i\bullet} = \textstyle\sum_{j=1}^{n} \pi_{ij} \quad \text{y} \quad \pi_{\bullet j} = \sum_{i=1}^{m} \pi_{ij}.$$

La probabilidad condicional $\mathbb{P}(Y = j \mid X = i)$ para cualesquiera i y j, se puede obtener de la siguiente manera, $\pi_{j|i} = \pi_{ij}/\pi_{i\bullet}$, $\forall$ i, j [1].

2.2.1. Sensibilidad y especificidad

Suponga que X es la variable que registra si un individuo tiene o no la enfermedad de Parkinson y que $\mathfrak{D}(X) = \{0, 1\}$, donde 1 equivale a "Sí" y 0 equivale a "No"; y suponga que Y es el resultado de una prueba para el diagnóstico de la enfermedad de Parkinson y que $\mathfrak{D}(Y) = \{0, 1\}$, donde 1 equivale a un resultado "Positivo" y 0 equivale a uno "Negativo".

Ejemplo de una tabla de contingencia de 2×2 con estructura de probabilidades.

	Diagnóstico		
Parkinson	Positivo	Negativo	Total
Sí	$\pi_{1\|1}$	$\pi_{0\|1}$	$\pi_{1\bullet}$
No	$\pi_{1\|0}$	$\pi_{0\|0}$	$\pi_{0\bullet}$
Total	$\pi_{\bullet 1}$	$\pi_{\bullet 0}$	$\pi_{\bullet\bullet}$

La tabla de contingencia anterior presenta las probabilidades de acierto y error de una cierta prueba de diagnóstico de la enfermedad de Parkinson. Dos de las probabilidades condicionales corresponden a un error en la prueba para el diagnóstico de la enfermedad, estas son $\pi_{0|1}$ y $\pi_{1|0}$. $\pi_{0|1}$ es la probabilidad de que el diagnóstico de la prueba sea negativo cuando el individuo tiene la enfermedad de Parkinson, $\pi_{1|0}$ es la probabilidad de que la prueba diagnostique al individuo con la enfermedad de Parkinson cuando no la tiene.

Por otro lado, las probabilidades condicionales restantes de la tabla de contingencia representan diagnósticos acertados. La *sensibilidad* (*sensitivity*) de la prueba es la probabilidad condicional de haber obtenido un diagnóstico positivo cuando se tiene la enfermedad ($\pi_{1|1}$), es la proporción de enfermos correctamente identificados; la *especificidad* (*specificity*) de la prueba es la probabilidad condicional de haber obtenido un diagnóstico negativo cuando no se tiene la enfermedad ($\pi_{0|0}$), es la proporción de sanos correctamente identificados.

En ocasiones es más común encontrar una tabla de contingencia que, en lugar de tener una estructura probabilística, presenta las frecuencias observadas en cada caso. En estas situaciones es posible estimar las probabilidades muestrales con base en la información disponible.

Ejemplo de una tabla de contingencia de 2×2 con estructura de frecuencias.

	Diagnóstico		
Parkinson	Positivo	Negativo	Total
Sí	n_{11}	n_{12}	$n_{1\bullet}$
No	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet} = n$

La estimación de la probabilidad condicional de la *ij-ésima* entrada de una tabla de contingencia de frecuencias es $\hat{\pi}_{j|i} = n_{ij}/n_{i\bullet}$ para $i, j \in \{1, 2\}$ [1].

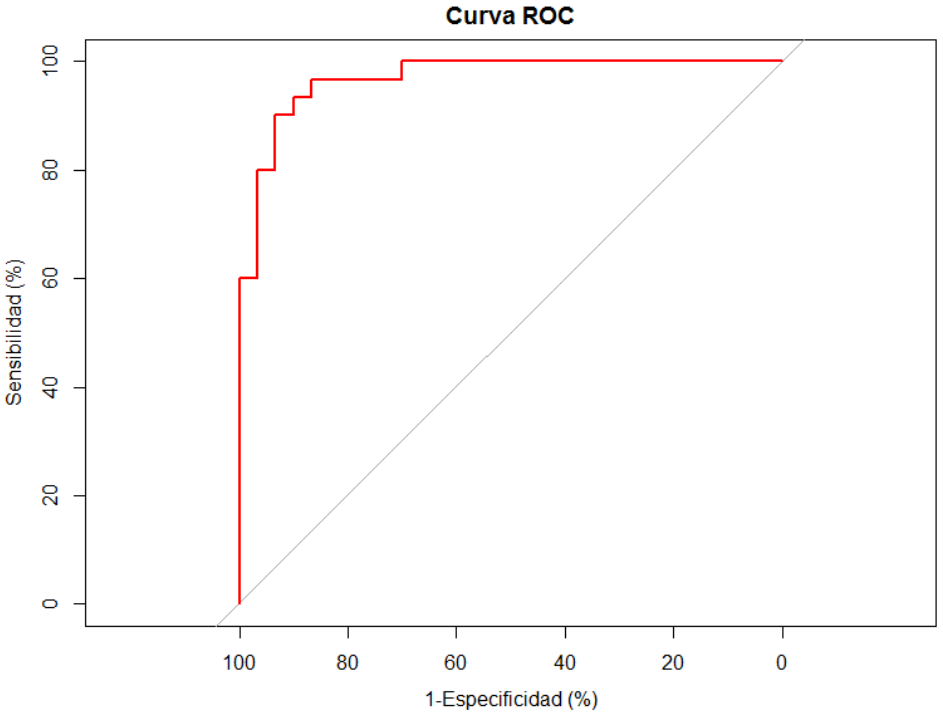
n_{11} es el número de *verdaderos positivos* (VP), n_{12} de *falsos negativos* (FN), n_{21} los *falsos positivos* (FP) y n_{22} los *verdaderos negativos* (VN). Las sumas por columna son los *positivos totales* (PT), $n_{\bullet 1}$; y $n_{\bullet 2}$ el número de *negativos totales* (NT).

2.2.2. Curva ROC

La sensibilidad y especificidad de una prueba dependen directamente del punto de corte que se considere para clasificar el resultado de una prueba como "Positivo" o "Negativo" [6]. Dada la naturaleza dicotómica de los resultados, es necesario utilizar el punto de corte como referencia para elegir alguna de las categorías cuando los resultados del método de diagnóstico que se haya propuesto arroje resultados decimales entre los valores cero y uno.

La curva de la *característica operativa del receptor* o *curva ROC* (por sus siglas en inglés, *Receiver Operating Characteristic*) es una gráfica de los valores (1-especificidad) contra la sensibilidad de la prueba para distintos puntos de corte [8].

El área bajo dicha curva representa la probabilidad de que un individuo con la enfermedad escogido aleatoriamente sea correctamente diagnosticado con mayor sospecha que un sujeto sin la enfermedad elegido aleatoriamente. El área bajo la curva ROC es un parámetro para cuantificar en un sólo valor numérico la situación general de una curva ROC con respecto a la diagonal no informativa, y tal área varía entre 0.5, en que se asegura que el método no tiene una precisión aparente, y 1, para un método que posee precisión aparente. La estimación más popular de la precisión de un método de diagnóstico es el área bajo una aproximación a la curva ROC ajustada [6].



Ejemplo de una curva ROC.
El área bajo la curva permite conocer la precisión del método de diagnóstico.

Si A_{ROC} representa el área bajo la curva ROC, la regla general para clasificar el método de diagnóstico es:

si $A_{ROC} = 0.5$,	no hay discriminación.
si $0.7 \leq A_{ROC} < 0.8$,	es considerada una discriminación aceptable.
si $0.8 \leq A_{ROC} < 0.9$,	es considerada una discriminación excelente.
si $A_{ROC} \geq 0.9$,	es considerada una discriminación sobresaliente.

Sin embargo, en la práctica es extremadamente inusual encontrar un $A_{ROC} \geq 0.9$, pues presupone que una separación completa es necesaria, lo cual haría imposible estimar los coeficientes de un modelo de regresión logística [8].

2.2.3. Riesgo relativo

Para el ejemplo planteado considere la simplificación de notación $\pi_{1|i} = \pi_i$ y del mismo modo para las probabilidades estimadas $\hat{\pi}_{1|i} = \hat{\pi}_i$.

π_i es la probabilidad de obtener un diagnóstico "Positivo" con la prueba, y $1 - \pi_i$ la de obtener un diagnóstico "Negativo", dado que la enfermedad de Parkinson está presente $(i = 1)$ o no $(i = 0)$.

Una medida importante en las investigaciones epidemiológicas y clínicas es el *riesgo relativo* (RR), el cual se define como

$$RR = \frac{\pi_1}{\pi_0}.$$

El RR por ser un cociente de probabilidades puede ser cualquier número real no negativo. Si $RR < 1$ la probabilidad de obtener un diagnóstico "Positivo", dado que no se tiene la enfermedad de Parkinson, es mayor que la de obtener el mismo diagnóstico teniendo la enfermedad. Por otro lado, si $RR = 1$ la probabilidad de obtener un resultado "Positivo" con la prueba de diagnóstico es la misma sin importar si se tiene o no la enfermedad. Ambas situaciones ($RR < 1$ o $RR = 1$) revelan que el método de diagnóstico en estudio no es eficiente para determinar correctamente la presencia de la enfermedad en el individuo. Si $RR > 1$ la probabilidad de obtener un resultado "Positivo", dado que la persona tiene la enfermedad de Parkinson, es mayor que la de obtener el mismo resultado sin tener la enfermedad. Este resultado ($RR > 1$) es un primer resultado favorable, que indica que el método de diagnóstico utilizado es eficiente para su propósito.

2.2.4. Razón de momios

El cociente $\Omega_i = \pi_i/(1 - \pi_i)$ representa las posibilidades de obtener un diagnóstico "Positivo" sobre uno "Negativo", dado que se tiene ($i = 1$) o no ($i = 0$) la enfermedad de Parkinson. Este cociente es ampliamente conocido como *momio*.

Considere el caso en que $i = 1$, es decir, que el individuo tiene la enfermedad de Parkinson, entonces si $\Omega_1 < 1$ hay $1/\Omega_1$ veces más posibilidad de ser diagnosticado como "Negativo" que como "Positivo"; si $\Omega_1 = 1$ se tienen las mismas posibilidades de obtener un diagnóstico "Positivo" y "Negativo"; si $\Omega_1 > 1$ hay Ω_1 veces más posibilidad de obtener el diagnóstico "Positivo" que el "Negativo". El análisis para el caso en que $i = 0$, i.e. el individuo no tiene la enfermedad de Parkinson, es análogo.

A pesar de la simpleza aritmética de estas operaciones no debe desestimarse el cálculo de este procedimiento, pues ofrece una perspectiva acerca de la efectividad del método de diagnóstico propuesto.

Otro concepto importante dentro del análisis de datos de una tabla de contingencia es la *razón de momios* (RM), la cual se define como

$$RM = \frac{\Omega_1}{\Omega_0} = \frac{\dfrac{\mathbb{P}(Y = 1 \mid X = 1)}{\mathbb{P}(Y = 0 \mid X = 1)}}{\dfrac{\mathbb{P}(Y = 1 \mid X = 0)}{\mathbb{P}(Y = 0 \mid X = 0)}} = \frac{\dfrac{\pi_1}{1 - \pi_1}}{\dfrac{\pi_0}{1 - \pi_0}}.$$

Por lo tanto RM puede tomar cualquier valor real no negativo. Si $RM < 1$ las posibilidades de tener como resultado de la prueba un diagnóstico "Positivo" sobre uno "Negativo" son mayores cuando no se tiene la enfermedad de Parkinson que cuando se tiene. Si $RM = 1$ las posibilidades del resultado del diagnóstico son iguales en presencia y en ausencia de la enfermedad, también se sigue que las probabilidades son independientes y que las probabilidades condicionales son iguales a las marginales. Si $RM > 1$ las posibilidades de tener como resultado de la prueba un diagnóstico "Positivo" sobre uno "Negativo" son mayores cuando se tiene la enfermedad de Parkinson que cuando no se tiene.

Un resultado inmediato es que la razón de posibilidades es un múltiplo del riesgo relativo

$$RM = RR \times \frac{1-\pi_0}{1-\pi_1}.$$

Dado que esperamos que exista una diferencia entre el diagnóstico de un individuo con la enfermedad de Parkinson y uno que no la tiene, i.e., esperamos que $RM \neq 1$, sería entonces conveniente encontrar el intervalo de confianza para la RM en el que podemos asegurar que existe una diferencia en el resultado del diagnóstico si se tiene la enfermedad. Para nuestra suerte existe algo similar, conocido como el intervalo de Wald.

Podemos estimar RM por medio de los datos disponibles en las frecuencias de la tabla de contingencia del siguiente modo

$$\widehat{RM} = \frac{\widehat{\Omega}_1}{\widehat{\Omega}_0} = \frac{\dfrac{\hat{\pi}_1}{1-\hat{\pi}_1}}{\dfrac{\hat{\pi}_0}{1-\hat{\pi}_0}} = \frac{\dfrac{\dfrac{n_{11}}{n_{11}+n_{12}}}{1-\dfrac{n_{11}}{n_{11}+n_{12}}}}{\dfrac{\dfrac{n_{21}}{n_{21}+n_{22}}}{1-\dfrac{n_{21}}{n_{21}+n_{22}}}} = \frac{\dfrac{n_{11}}{n_{12}}}{\dfrac{n_{21}}{n_{22}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

El estimador $\widehat{RM}$ posee una distribución normal asintótica alrededor de RM. Sin embargo, se utiliza la transformación $\log \widehat{RM}$ porque la estructura aditiva del logaritmo converge más rápidamente a una distribución normal que la estructura multiplicativa de RM. El estimador de la varianza para $\log \widehat{RM}$ es

$$\mathbb{V}\left(\log\left(\widehat{RM}\right)\right) \approx \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$ [1]

[1] La obtención de este resultado se encuentra en la sección de Anexos.

Entonces el intervalo de Wald con $(1-\alpha)\times 100\,\%$ de confianza para $\log(RM)$ es

$$\log\left(\widehat{RM}\right) \pm z_{\alpha/2}\sqrt{\mathbb{V}\left(\log\left(\widehat{RM}\right)\right)},$$

donde $z_{\alpha/2}$ es el cuantil $\alpha/2$ de una distribución normal estándar [1].

Por lo tanto, el intervalo con $(1-\alpha)\times 100\,\%$ de confianza para RM es

$$\exp\left\{\log\left(\widehat{RM}\right) \pm z_{\alpha/2}\sqrt{\mathbb{V}\left(\log\left(\widehat{RM}\right)\right)}\right\}.$$

Con base en esto, es posible asegurar que RM es estadísticamente distinto de uno al nivel de significancia α si el intervalo de confianza anterior no incluye al uno.

Capítulo 3

Análisis de componentes principales

Considere un vector p-variado $X^T = (X_1, X_2, \ldots, X_p)$, donde cada X_i es una variable aleatoria asociada a una cierta característica de un individuo, $1 \leq i \leq p$. Existen diversas maneras de hacer que las p variables correlacionadas se transformen en m variables no correlacionadas, $m \leq p$, una de ellas es la transformación lineal conocida como *componentes principales*. Suponga además que X tiene segundo momento finito y esperanza cero, entonces es posible obtener su matriz de covarianzas. La matriz de covarianzas mide la dispersión de los datos de una variable (varianzas) y las relaciones entre dos variables (covarianzas), y son utilizadas para comparar modelos que contienen información provista por tales variables [3].

Suponga que se tienen k matrices de covarianzas, $k > 1$, $k \in \mathbb{N}$, entonces es posible decir que, en general, es una mala práctica comparar un modelo en que se consideran las k matrices de covarianzas idénticas contra uno en que se consideran distintas, debido a la diferencia tanto de los grados de libertad entre los modelos como del número de parámetros [4]. Por ello existen distintos modelos o niveles de similaridad que proveen información acerca de la naturaleza de las diferencias entre matrices de covarianzas sin agregar tantos grados de libertad, uno de ellos es el *análisis de componentes principales* (ACP) que compara matrices de covarianza de acuerdo a sus componentes principales.

Al multiplicar una *matriz ortogonal*[1] $\mathbf{A}$ por un vector X se obtiene una *transformación ortogonal* $\mathrm{T}(X) = \mathbf{A}X$ que preserva normas, ángulos y distancias. Estas características son muy importantes para el análisis de componentes principales, porque aseguran que los componentes escogidos en los espacios ortogonales preservan la estructura de las variaciones del espacio original no rotado.

[1] Véase en el Apéndice A.

El análisis de componentes principales echa mano de herramientas matemáticas como la *descomposición espectral* de una matriz, a fin de reducir la dimensión de los datos, creando una *base ortogonal* del subespacio real $\mathbb{R}^m \subseteq \mathbb{R}^p$, tal que $\mathbb{R}^m$ sea la mejor reconstrucción de $\mathbb{R}^p$.

Sea $\Psi = \mathbb{E}(XX^T) \in \mathbb{M}_{p\times p}$ la matriz de covarianzas de X, sin pérdida de generalidad puede escogerse una matriz $\beta = (\beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{M}_{p\times p}$ cuyos vectores columna sean base del espacio $\mathbb{R}^p$, entonces por el teorema de descomposición espectral

$$\mathbf{\Psi} = \beta \mathbf{\Lambda} \beta^T, \tag{3.1}$$

donde β es una matriz ortogonal y $\mathbf{\Lambda}$ es la *matriz diagonal*

$$\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$

Cada uno de los λ_i recibe el nombre de *eigenvalor*, *raíz característica* o *valor característico*, $i = 1, \ldots, p$. El método de componentes principales asocia la i-ésima raíz característica al i-ésimo componente y posteriormente selecciona aquellos que conservan la mayor variabilidad, y dado que la matriz $\mathbf{\Lambda}$ contiene los valores característicos de la matriz de covarianzas $\mathbf{\Psi}$, sin pérdida de generalidad, suponga que los valores de las raíces características λ_i se encuentran ordenados de forma no creciente, i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ [4].

Los p valores característicos, $\lambda_1, \lambda_2, \ldots, \lambda_p$, se obtienen de las raíces del polinomio de grado p, $det\,(\mathbf{\Psi} - \lambda \mathbb{I}_p) = 0$, donde la expresión "*det*" se refiere al *determinante* de la matriz $\mathbf{\Psi} - \lambda \mathbb{I}_p$.

3.1. Componentes principales

El método de componentes principles consiste en descomponer la matriz de covarianzas en las direcciones en que se encuentra la variación de los datos. Lo cual se logra al encontrar una transformación ortogonal ordenada, en la que la primer componente se asocia con el eje de mayor variación de los datos y la última con el eje en que haya menor variación, tal que el i-ésimo componente obtenido no se encuentre correlacionado con los $i-1$ componentes anteriores, $i = 2, \ldots, p$ [3].

Sea $\mathrm{T}(X) = \mathbf{A}^T X$, entonces

$$\begin{aligned}
T_1 &= a_1^T X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\
T_2 &= a_2^T X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\
&\vdots \\
T_j &= a_j^T X = a_{j1}X_1 + a_{j2}X_2 + \ldots + a_{jp}X_p \\
&\vdots \\
T_p &= a_p^T X = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p
\end{aligned}$$

y matriz de covarianzas

$$\begin{aligned}
\mathbf{\Psi}_{T(X)} &= \mathbb{V}\left[\ T(X)\ \right] = \mathbb{E}\left[\ \{\mathbf{A}^T X\}\ \{\mathbf{A}^T X\}^T\ \right] \\
&= \mathbb{E}\left[\ \mathbf{A}^T X X^T \mathbf{A}\ \right] = \mathbf{A}^T\ \mathbb{E}\left[\ X X^T\ \right]\ \mathbf{A} \\
&= \mathbf{A}^T \mathbf{\Psi} \mathbf{A},
\end{aligned}$$

de la ortogonalidad de $\mathbf{A}$ y $\mathbf{\Psi}$ se sigue que $\mathbf{\Psi}_{T(X)}$ es ortogonal.

Dado el vector X buscamos cómo extraer los k componentes principales, $k \leq p$, tales que $a_j^T a_j = 1$, $1 \leq j \leq p$, y $\mathbf{\Psi}_{T_1} \geq \mathbf{\Psi}_{T_2} \geq \ldots \geq \mathbf{\Psi}_{T_p}$. La primera condición se satisface debido a que $\mathbf{A}$ es una matriz ortogonal y los elementos de la diagonal de $\mathbf{A}^T\mathbf{A}$ son los mismos que los de la matriz identidad $\mathbb{I}_p$.

El primer componente principal es $T_1 = a_1^T X$, sujeto a $a_1^T a_1 = 1$, tal que a_1 maximiza a $\mathbf{\Psi}_{T_1}$. Sucesivamente se buscan los componentes que cumplen un conjunto cada vez mayor de restricciones de manera que el i-ésimo componente no se encuentre correlacionado con los anteriores. El segundo componente principal es $T_2 = a_2^T X$, sujeto a $a_2^T a_2 = 1$, tal que a_2 maximiza a $\mathbf{\Psi}_{T_2}$ y $Cov(T_1, T_2) = 0$. El k-ésimo componente principal es $T_k = a_k^T X$, sujeto a $a_k^T a_k = 1$, tal que a_k maximiza a $\mathbf{\Psi}_{T_k}$ y $Cov(T_j, T_k) = 0$, $\forall\ j < k$.

El objetivo general es encontrar una matriz $\mathbf{A}$ que maximice $\mathbf{A}^T\mathbf{\Psi}\mathbf{A}$, tal que $\mathbf{A}$ sea una matriz ortogonal. Haciendo uso de los *multiplicadores de Lagrange*, el problema puede resolverse maximizando la función

$$\mathbf{F} = \mathbf{A}^T\mathbf{\Psi}\mathbf{A} - \mathbf{\Lambda}(\mathbf{A}^T\mathbf{A} - \mathbb{I}_p),$$

lo cual se obtiene al diferenciar la función $\mathbf{F}$ con respecto a la matriz $\mathbf{A}$ e igualarlo a la matriz cero de $\mathbb{M}_{p\times p}$.

$$\frac{\partial \mathbf{F}}{\partial \mathbf{A}} = 2\mathbf{A}^T\mathbf{\Psi} - 2\mathbf{\Lambda}\mathbf{A}^T = \mathbf{0}$$

Dado que $\mathbf{A}$ es una matriz ortogonal, es equivalente multiplicar por $\mathbf{A}^T$ por la izquierda que por la matriz $\mathbf{A}$ por la derecha. Por lo tanto, $\mathbf{A}^T\mathbf{\Psi} - \mathbf{\Lambda}\mathbf{A}^T = \mathbf{0}$ o equivalentemente $\mathbf{\Psi}\mathbf{A} - \mathbf{A}\mathbf{\Lambda} = \mathbf{0}$. Entonces $\mathbf{A}^T\mathbf{\Psi} = \mathbf{\Lambda}\mathbf{A}^T$, de manera que

$$\mathbf{\Psi} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T,$$

i.e., el problema se reduce a encontrar la descomposición espectral de $\mathbf{\Psi}$ como en (3.1).

Una vez que se obtiene la matriz $\mathbf{\Lambda}$, es posible encontrar la matriz $\mathbf{A}$ a través del sistema de ecuaciones $(\mathbf{\Psi} - \mathbf{\Lambda})\mathbf{A} = \mathbf{0}$. A partir de esto se sustituyen los valores de las columnas de la matriz $\mathbf{A}$ para cada uno de los componentes, desde T_1 hasta T_p, y se seleccionan los primeros k componentes, $k \leq p$, rechazando los $p - k$ restantes cuya variabilidad no es significativa.

El criterio de selección de los primeros k componentes, $k = 1, \ldots, p$, en cada caso, depende del nivel de variabilidad que desea conservar para el grupo de componentes que se está analizando. Suponga que desea incluir los componentes que conservan el $(\omega \times 100)\,\%$ de variabilidad, $\omega \in [0, 1]$, entonces se realiza el siguiente cociente

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j},$$

donde el numerador corresponde a la suma de los primeros k valores propios, tal que el cociente $\left(\sum_{j=1}^{k-1} \lambda_j / \sum_{j=1}^{p} \lambda_j\right) < (\omega \times 100)\,\%$ y $\left(\sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{p} \lambda_j\right) \geq (\omega \times 100)\,\%$.

Un criterio importante para la comprobación del algoritmo es el siguiente

$$tr\left(\mathbf{\Psi}_{T(X)}\right) = \sum_{j=1}^{p} \lambda_j,$$

donde "*tr*" se refiere a la *traza* de la matriz $\mathbf{\Psi}_{T(X)}$ y los λ_j's son las raíces o valores característicos de la matriz diagonal $\mathbf{\Lambda}$ [4].

Otro criterio importante [4], aunque más difícil de comprobar, especialmente para dimensiones grandes de la matriz de covarianzas, es el siguiente

$$det\left(\mathbf{\Psi}_{T(X)}\right) = \prod_{j=1}^{p} \lambda_j. \tag{3.2}$$

3.1.1. Componentes principales de una muestra

Cuando se cuenta con una base de datos numéricos, la teoría de componentes principales ofrece una solución para la obtención de dichos componentes, cuya justificación se

encuentra en los comportamientos asintóticos teóricos.

Suponga que se tiene una muestra aleatoria de tamaño N de vectores p-variados, a saber, $X_1, X_2, \ldots, X_N$, entonces se define el vector de la media muestral como

$$\overline{X} = \frac{1}{N} \sum_{j=1}^{N} X_j$$

y la matriz de covarianzas muestral como

$$\mathbf{S} = \frac{1}{N-1} \sum_{j=1}^{N} \left(X_j - \overline{X}\right) \left(X_j - \overline{X}\right)^T. \tag{3.3}$$

Al igual que en el caso teórico, conocer la descomposición espectral de la matriz de covarianzas $\mathbf{S}$ es de vital importancia para encontrar los componentes principales muestrales. Por el teorema de descomposición espectral

$$\mathbf{S} = \mathbf{BLB}^T,$$

donde $\mathbf{B}$ resulta ser el estimador máximo verosímil de β, i.e. $\mathbf{B} = \widehat{\beta}$, y la matriz diagonal $\mathbf{L}$ el estimador máximo verosímil de $\mathbf{\Lambda}$, i.e. $\mathbf{L} = \widehat{\Lambda}$, con β y $\mathbf{\Lambda}$ como en (3.1). Una vez que se tiene la matriz $\mathbf{B}$, lo siguiente es obtener los componentes principales de la misma manera que en el caso teórico [4].

La mayoría de las aplicaciones prácticas del ACP están basadas en matrices de correlaciones, más que en matrices de covarianzas, especialmente en aplicaciones en que las unidades de medida son arbitrarias (o no unificadas). Sin embargo, en el caso multivariado, los estimadores obtenidos a partir de las matrices de correlaciones no necesariamente son los estimadores máximo verosímiles y la distribución asintótica de los componentes principales y las raíces características de la matriz son difíciles de obtener [4].

3.2. Componentes principales comunes

En ocasiones, una muestra puede ser dividida en grupos, e.g. en un grupo control y un grupo con individuos que poseen cierta característica no presente en el grupo control, para los propósitos de un cierto estudio, de modo que el interés está situado en encontrar los componentes que mejor representen la variabilidad de los datos y que sean comunes a ambos grupos o submuestras. Bajo la ocurrencia de tales circunstancias, el *análisis de componentes principales comunes* (análisis de *CPC* o *ACPC*) es especialmente útil para obtener los componentes principales de cada uno de los grupos.

El ACPC es un modelo que estima la semejanza entre matrices de covarianzas en base a si es posible encontrar una base ortogonal β común a todas ellas. Esta técnica sirve para formalizar que todos los grupos o poblaciones considerados poseen los mismos componentes principales [4].

Suponga que los valores observados de los individuos de la muestra están asociados a J grupos mutuamente excluyentes, i.e. que si los datos de un individuo pertenecen al grupo h entonces no se encuentran en el grupo i, para $h \neq i$, tal que cada uno de los grupos tiene matriz de covarianzas $\mathbf{\Psi}_i$, $\mathbf{\Psi}_h \in \mathbb{M}_{p\times p}$ definida positiva, $h, i = 1, \ldots, J$, donde p es el número de variables en los grupos.

El modelo de componentes principales comunes es el siguiente

$$\mathbf{\Psi}_i = \beta \mathbf{\Lambda}_i \beta^T, \tag{3.4}$$

para $i = 1, \ldots, J$, donde $\beta \in \mathbb{M}_{p\times p}$ es una matriz ortogonal y $\mathbf{\Lambda}_i$ es la matriz diagonal

$$\mathbf{\Lambda}_i = diag(\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ip}) = \begin{bmatrix} \lambda_{i1} & 0 & \cdots & 0 \\ 0 & \lambda_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{ip} \end{bmatrix}. \tag{3.5}$$

El número de parámetros en el modelo de CPC es $p(p-1)/2$ para la matriz β y $J \cdot p$ para las matrices diagonales [4].

De este modo es posible seleccionar los CPC a los J grupos existentes a partir de la matriz de vectores propios $\mathbf{\Lambda}_i$, con $i = 1, \ldots, J$, para el i-ésimo grupo. Los componentes principales comunes se seleccionan de manera similar que en el ACP, dependiendo del porcentaje de variabilidad que se desea conservar.

3.2.1. Estimación máximo verosímil del modelo de CPC

Para saber si es posible rechazar o no la hipótesis de CPC, es decir, si es posible o no formar categorías o grupos entre las variables de modo que para todos ellos pueda descomponerse espectralmente su matriz de covarianzas bajo la misma base ortogonal β, y asegurarnos de tomar de entre todos los componentes aquellos que mejor explican el comportamiento del modelo, existe la siguiente prueba de hipótesis del modelo de CPC

$$H_{CPC} : \mathbf{\Psi}_i = \beta \mathbf{\Lambda}_i \beta^T, \qquad i = 1, \ldots, J,$$

y requiere que las matrices $\mathbf{\Lambda}_i$ se encuentren bien definidas, i.e. que exista al menos una i, con $i \neq k$, tal que $\lambda_{il} \neq \lambda_{kl}$, para $i, k \in \{1, \ldots, J\}$, $\forall\ l \in \{1, \ldots, p\}$, y que la matriz de covarianzas $\mathbf{\Psi}_i$ sea positiva definida $\forall\ i \in \{1, \ldots, J\}$ [4].

Sea N_i el tamaño de cada grupo, $i = 1, \ldots, J$. Entonces es posible obtener las matrices de covarianzas muestrales de cada grupo, $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_J$, como en (3.3), pero sumando sólo los N_i valores del i-ésimo grupo para obtener la matriz de covarianzas $\mathbf{S}_i$.

De este modo, la función de verosimilitud conjunta de $\mathbf{\Psi}_1, \mathbf{\Psi}_2, \ldots, \mathbf{\Psi}_J$ dadas las matrices $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_J$ es

$$\mathbf{L}\left(\mathbf{\Psi}_1, \mathbf{\Psi}_2, \ldots, \mathbf{\Psi}_J\right) = C \times \prod_{i=1}^{J} \exp\left\{-\frac{N_i - 1}{2}\ tr\left(\mathbf{\Psi}_i^{-1}\mathbf{S}_i\right)\right\} \left(det\ \mathbf{\Psi}_i\right)^{-(N_i-1)/2},$$

donde C es una cantidad que no depende de los parámetros [4].

Suponiendo que el modelo de CPC se cumple, la función de verosimilitud alcanza su máximo cuando $\mathbf{\Psi}_i = \mathbf{S}_i$, $i = 1, 2, \ldots, J$. Por lo tanto,

$$tr\left(\mathbf{\Psi}_i^{-1}\mathbf{S}_i\right) = tr\left(\beta\mathbf{\Lambda}_i^{-1}\beta^T\mathbf{S}_i\right) = tr\left(\mathbf{\Lambda}_i^{-1}\beta^T\mathbf{S}_i\beta\right) = tr\left(\mathbf{\Lambda}_i^{-1}\mathbf{\Lambda}_i\right) = tr\left(\mathbb{I}_p\right) = p.$$

Y el estimador máximo verosímil de $\mathbf{\Psi}_i$ es

$$\widehat{\mathbf{\Psi}}_i = \widehat{\beta}\ \widehat{\mathbf{\Lambda}}_i\ \widehat{\beta}^{\,T}, \qquad i = 1, 2, \ldots, J.$$

El estadístico de prueba de la hipótesis de CPC es el cociente de verosimilitud

$$\begin{aligned}
X^2_{CPC} &= -2\ \log\ \frac{\mathbf{L}(\widehat{\mathbf{\Psi}}_1, \widehat{\mathbf{\Psi}}_2, \ldots, \widehat{\mathbf{\Psi}}_J)}{\mathbf{L}\left(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_J\right)} \\
&= -2\ \log\ \frac{C \times \prod_{i=1}^{J} \exp\left\{-\frac{N_i-1}{2}\ p\right\} \left(det\ \widehat{\mathbf{\Psi}}_i\right)^{-(N_i-1)/2}}{C \times \prod_{i=1}^{J} \exp\left\{-\frac{N_i-1}{2}\ p\right\} \left(det\ \mathbf{S}_i\right)^{-(N_i-1)/2}} \\
&= -2\ \log\ \prod_{i=1}^{J} \left(\frac{det\ \widehat{\mathbf{\Psi}}_i}{det\ \mathbf{S}_i}\right)^{-(N_i-1)/2} \\
&= -2 \sum_{i=1}^{J}\ \log \left(\frac{det\ \widehat{\mathbf{\Psi}}_i}{det\ \mathbf{S}_i}\right)^{-(N_i-1)/2} \\
&= \sum_{i=1}^{J}\ (N_i - 1)\ \log \left(\frac{det\ \widehat{\mathbf{\Psi}}_i}{det\ \mathbf{S}_i}\right),
\end{aligned}$$

donde $X^2_{CPC} \sim \chi^2_{((J-1)p(p-1)/2)}$ bajo H_{CPC}. Se rechaza H_{CPC} a un nivel de significancia α si $X^2_{CPC} > \chi^2_{((J-1)p(p-1)/2,\ 1-\alpha)}$, el cuantil $1-\alpha$ de una distribución ji-cuadrada con $(J-1)p(p-1)/2$ grados de libertad [4].

Capítulo 4

Regresión logística

4.1. Modelos lineales generalizados

Considere un espacio muestral $\mathcal{S}$, un espacio paramétrico Θ y una función P, tal que $P : \Theta \to \mathcal{P}(\mathcal{S})$, donde $\mathcal{P}(\mathcal{S})$ es el conjunto de todas las distribuciones de probabilidad sobre el espacio muestral $\mathcal{S}$. Entonces P es una función que a cada $\theta \in \Theta$ le asigna una función de distribución de probabilidad en $\mathcal{P}(\mathcal{S})$ [10].

Un *modelo estadístico* P es la especificación de una distribución de probabilidad $P(\theta)$, una vez que se conoce el parámetro θ. Sea $X^T = (X_1, X_2, \ldots, X_p)$ un vector aleatorio p-variado. Las combinaciones lineales de las variables X_i's, $1 \leq i \leq p$, pueden determinar el comportamiento esperado de la distribución de probabilidad $P(\theta)$ de la *variable respuesta*. Es precisamente por este hecho que se conoce a las covariables X_i's, $1 \leq i \leq p$, como *variables explicativas* o *regresoras*.

Se considera un hecho favorable cuando la variable respuesta toma una distribución $P(\theta)$ conocida, cuando una combinación lineal de las variables regresoras determina cabalmente el comportamiento esperado de esta respuesta. Algunos modelos con este tipo de conducta se conocen como *modelos lineales generalizados*.

4.1.1. Familia de dispersión exponencial

Considere el vector respuesta Y, cuyas entradas corresponden a los valores observados y_i de la variable respuesta, y_i es independiente de y_j para $i \neq j$, $1 \leq i, j \leq n$, producto de la aleatoriedad del proceso. El componente aleatorio de un modelo lineal generalizado consiste de una variable respuesta Y con observaciones independientes de una distribución en la familia exponencial natural. La familia exponencial tiene función de densidad

de probabilidad o función de masa de la forma

$$f(y_i;\theta,\phi) = \exp\left\{\left[y_i\theta - b(\theta)\right]/a(\phi) + c(y_i,\phi)\right\}.$$

$f(y_i;\theta,\phi)$ se conoce como la *familia de dispersión exponencial*, ϕ como el *parámetro de dispersión* y θ como el *parámetro natural*. Las variables aleatorias cuya función de masa o densidad de probabilidad puede escribirse de esta forma, pertenecen a esta familia [1].

Un ejemplo de distribución en la familia de dispersión exponencial es la distribución normal [10]:

Si $Y \sim N(\mu,\sigma^2)$, entonces

$$\begin{aligned} f_Y(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\dfrac{(y-\mu)^2}{2\sigma^2}} = \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right\} \\ &= \exp\left\{-\frac{y^2-2\mu y+\mu^2}{2\sigma^2} + \log\left(\sqrt{2\pi\sigma^2}^{\,-1}\right)\right\} \\ &= \exp\left\{\frac{2\mu y-\mu^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right) - \frac{y^2}{2\sigma^2}\right\} \end{aligned}$$

Sea $\theta = \mu$, $b(\theta) = \theta^2/2$, $\phi = \sigma$ y $a(\phi) = \sigma^2$,

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} - \log\left(\sqrt{2\pi\phi^2}\right) - \frac{y^2}{2\phi^2}\right\}.$$

Si definimos $c(y,\phi) = -(\log(\sqrt{2\pi\phi^2}\,) + y^2/2\phi^2)$, entonces

$$f_Y(y;\theta,\phi) = \exp\left\{\left[y\theta - b(\theta)\right]/a(\phi) + c(y,\phi)\right\}.$$

Otros ejemplos de distribuciones pertenecientes a esta familia son la distribución Gaussiana inversa y la gamma.

4.1.1.1. Familia exponencial natural

Cuando ϕ y $a(\phi)$ son conocidos, la familia $f(y_i;\theta,\phi)$ puede escribirse como $f(y_i;\theta)$, la cual recibe el nombre de *familia exponencial natural*, porque las distribuciones de esta familia quedan totalmente definidas por el parámetro natural θ.

Si $a_0(\theta) = \exp\left[-\dfrac{b(\theta)}{a(\phi)}\right]$, $b_0(y_i) = \exp\left[c(y_i, \phi)\right]$ y $Q(\theta) = \dfrac{\theta}{a(\phi)}$, entonces

$$\begin{aligned} f(y_i; \theta, \phi) &= \exp\left\{ \frac{[y_i\theta - b(\theta)]}{a(\phi)} + c(y_i, \phi) \right\} \\ &= a_0(\theta)\ b_0(y_i) \exp\left[y_i\ Q(\theta)\right] = f(y_i; \theta), \end{aligned}$$

a $Q(\theta)$ se le llama *parámetro natural*.

Por otro lado, se conoce como *liga canónica* a la *función liga* o *función enlace* que asocia la media de la distribución $f(y; \theta)$ con el parámetro natural $Q(\theta)$.

Un ejemplo de distribución en la familia exponencial natural es la distribución Bernoulli:

Si $Y \sim Bernoulli(\pi)$, entonces

$$\begin{aligned} f(y; \pi) &= \pi^y\ (1-\pi)^{1-y} = (1-\pi)\left[\pi/(1-\pi)\right]^y \\ &= (1-\pi) \exp\left[y\ \log\left(\frac{\pi}{1-\pi}\right)\right], \end{aligned}$$

donde $a_0(\pi) = 1 - \pi$, $b_0(y) = 1$ y $Q(\pi) = \log\left[\pi/(1-\pi)\right]$. La función $\log\left[\pi/(1-\pi)\right]$ es la función que asocia a la media de Y, $\mathbb{E}[Y] = \pi$, con el parámetro natural, por lo tanto $Q(\pi)$ es la liga canónica de una variable aleatoria con distribución Bernoulli(π). Otras distribuciones de esta familia son la distribución binomial y la Poisson.[1]

4.1.2. Modelos de respuesta binaria

En la teoría existen distintos modelos lineales generalizados que permiten analizar los datos disponibles de modo que sea posible realizar predicciones acerca de la respuesta que inducen los mismos.

El estudio de una variable Y de respuesta binaria ($\{0, 1\}$) sujeta a la información de p variables regresoras $X^T = (X_1, X_2, \ldots, X_p)$ puede modelarse a través de una distribución Bernoulli($\pi(x)$), donde $\pi(x)$ representa la probabilidad de que la variable respuesta sea igual a 1, dado que X toma los valores $x^T = (x_1, x_2, \ldots, x_p)$, i.e. $\mathbb{P}(Y = 1 \mid X = x)$ $= \pi(x)$.

La liga canónica asociada a una variable aleatoria Bernoulli es conocida como *logit*,

$$logit\ (\pi(X)) = \log\left(\frac{\pi(X)}{1-\pi(X)}\right).$$

[1] Algunas distribuciones de la familia exponencial natural se encuentran en el cuadro A.1 en el Apéndice A.

Una variable aleatoria binomial con parámetros n y $\pi(X)$, posee la misma función liga *logit*, este resultado se obtiene al factorizar su función de masa de probabilidad en los componentes de la familia exponencial natural. La tercera y última característica de un modelo lineal generalizado, además de poseer una función liga asociada a la media y pertenecer a la familia exponencial natural, es que la función liga se encuentra relacionada con el modelo lineal, de este modo

$$logit\ (\pi(X)) = \beta^T X.$$

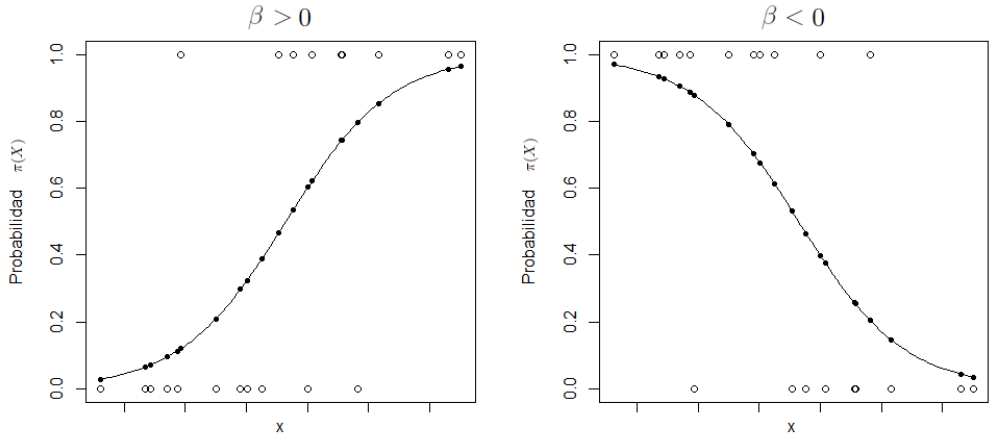
De manera equivalente, esta relación se puede expresar como

$$\pi(X) = \frac{exp(\beta^T X)}{1 + exp(\beta^T X)}, \tag{4.1}$$

donde β es el vector de parámetros de regresión $\beta^T = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ y $X^T = (1, X_1, X_2, \ldots, X_p)$.

$\pi(X)$ definida en (4.1) se conoce como el *modelo de regresión logística*. Existen otras funciones liga para modelar la relación entre una respuesta binaria y un conjunto de variables regresoras, las más conocidas son las ligas *probit*, *loglog* y *cloglog*.[2] La elección de la liga depende del problema que se desea resolver.

La siguiente figura muestra la relación entre una variable explicativa X y la probabilidad $\pi(X)$, para el caso en que el parámetro asociado a X sea positivo o negativo, es decir, cuando $\beta > 0$ o $\beta < 0$.



Gráfica del modelo de regresión logística para el estudio de una variable de respuesta binaria. Los puntos por debajo y por encima de la curva representan probabilidades 0 y 1, respectivamente.

[2] Estas funciones liga se enuncian en el cuadro A.2 en el Apéndice A.

4.2. Ajuste del modelo

La *función de verosimilitud* expresa la probabilidad de observar los valores de la respuesta de un evento, dado que se tienen los parámetros β y los valores $(x_1, x_2, \ldots, x_p)$. En este sentido, la función de verosimilitud puede verse como una *función objetivo*, para la cual deseamos estimar los parámetros de β que maximicen la probabilidad de ocurrencia de las respuestas observadas con los valores observados de X. Cuando se han encontrado estos parámetros, dicha función recibe el nombre de *función de máxima verosimilitud*.

Sean ${x_i}^T = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ los valores observados pertenecientes a la muestra del i-ésimo individuo, x_i se conoce como un *conjunto de covariables* y se dice que los individuos i y j poseen el mismo conjunto de covariables si $x_i = x_j$, $i, j = 1, \ldots, n$. Del mismo modo, $\pi(x_i)$ representa la probabilidad de que la respuesta sea positiva, dado que se tiene el conjunto de covariables x_i y es obtenida como en (4.1).

Suponga que se tienen n observaciones independientes (x_i, y_i), $i = 1, 2, \ldots, n$, tales que x_i y y_i son el conjunto de covariables y el valor observado de la respuesta del i-ésimo individuo, respectivamente. Entonces la función de verosimilitud conjunta de las n observaciones es:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}. \tag{4.2}$$

La naturaleza multiplicativa de $l(\beta)$ puede complicar la estimación de los parámetros de β, por ello se utiliza la función de *log*-verosimilitud, definida como $L(\beta) = \log(l(\beta))$,

$$L(\beta) = \sum_{i=1}^{n} y_i \log[\pi(x_i)] + (1 - y_i) \log[1 - \pi(x_i)]. \tag{4.3}$$

Esta función alcanza su valor máximo en los mismos valores que la función de verosimilitud por tratarse de una transformación biyectiva monótona creciente, además, su cualidad aditiva facilita los cálculos para encontrar los parámetros desconocidos. Derivando la expresión $L(\beta)$ con respecto de cada uno de los parámetros buscados e igualando a cero, se obtiene el siguiente sistema de *p+1* ecuaciones de máxima verosimilitud:

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0,$$
$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0,$$

para $j = 1, \ldots, p$. Donde los valores estimados $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ son las soluciones del sistema [8].

Es necesario que, una vez determinados los parámetros, podamos identificar cuáles de ellos son significativos, y cuáles son estadísticamente iguales a cero y no aportan información al modelo, para ello, además de realizar pruebas de hipótesis, estimaremos sus intervalos de confianza.

La teoría de estimación por el método de máxima verosimilitud asegura que es posible obtener los errores estándar de los coeficientes estimados por medio de invertir la *matriz de información observada de Fisher*, denotada como $\mathbf{I}(\beta)$, cuyas entradas son:

$$\mathbf{I}(\beta)_{jj} = \frac{\partial^2 L(\beta)}{\partial {\beta_j}^2} = -\sum_{i=1}^{n} {x_{ij}}^2 \pi_i (1 - \pi_i),$$
$$\mathbf{I}(\beta)_{jl} = \frac{\partial^2 L(\beta)}{\partial \beta_j \, \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i),$$

para $0 \leq j,\ l \leq p$, $x_{i0} = 1$ y $\pi_i = \pi(x_i)$ [8].

En la práctica, la matriz $\mathbf{I}(\beta)$ puede estimarse por medio de la descomposición $\widehat{\mathbf{I}}(\hat{\beta}) = \mathbf{X}^T \mathbf{V} \mathbf{X}$, donde $\mathbf{X} \in \mathbb{M}_{n \times (p+1)}$, $\mathbf{V} \in \mathbb{M}_{n \times n}$, y

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1(1 - \hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1 - \hat{\pi}_2) & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1 - \hat{\pi}_n) \end{bmatrix},$$

donde $\hat{\pi}_i$ representa la estimación de $\pi(x_i)$ utilizando $\hat{\beta}$.

Entonces es posible obtener la matriz de varianzas $\mathbb{V}(\beta) = \mathbf{I}^{-1}(\beta)$ por medio de $\widehat{\mathbf{I}^{-1}}(\hat{\beta})$. Naturalmente, $\widehat{\mathbb{V}}(\hat{\beta}_j)$ es la varianza estimada del j-ésimo valor estimado de $\hat{\beta}$ y por ello el error estándar de $\hat{\beta}_j$ es $\hat{v}[\hat{\beta}_j] = \left[\widehat{\mathbb{V}}(\hat{\beta}_j)\right]^{1/2}$.

Por lo tanto, el intervalo de $(1 - \alpha) \times 100\,\%$ de confianza para β_j es

$$\hat{\beta}_j \pm z_{\alpha/2} \cdot \hat{v}[\hat{\beta}_j],$$

donde $z_{\alpha/2}$ es el cuantil $\alpha/2$ de una distribución normal estándar [8].

Aseguramos que β_j es estadísticamente distinto de cero al nivel de significancia α si el intervalo anterior no incluye al cero.

4.2.1. Prueba del cociente de verosimilitudes

Recordemos que $\hat{\beta}$ permite obtener la máxima log-verosimilitud y verosimilitud, simultáneamente, de este modo, el modelo de regresión logística ajustado, con función de log-verosimilitud $L(\hat{\beta})$, recibe el nombre de *modelo saturado*; mientras que el modelo en el cual no existe influencia alguna de las variables, i.e. $\beta = \underline{\mathbf{0}}$, es conocido como el *modelo nulo*. La *prueba del cociente de verosimilitudes* mide la proporción en que se modifica la probabilidad de obtener los datos observados si ninguna de las variables regresoras tuviera un impacto significativo en la respuesta, es en este sentido que la prueba se centra en observar el comportamiento en relación a $\beta = \underline{\mathbf{0}}$:

$$H_0\colon \beta = \underline{\mathbf{0}} \qquad \textit{versus} \qquad H_1\colon \beta \neq \underline{\mathbf{0}}.$$

El estadístico de prueba es

$$\mathrm{G} = -2\log\left[\frac{l(\beta)}{l(\hat{\beta})}\right] = -2\left[L(\beta) - L(\hat{\beta})\right],$$

donde $\beta = \underline{0}$, y G tiene una distribución asintótica ji-cuadrada con p grados de libertad bajo H_0. Se rechaza H_0 a un nivel de significancia α si $\mathrm{G} > \chi^2_{(p,\ 1-\alpha)}$, con $\chi^2_{(p,\ 1-\alpha)}$ el cuantil $1-\alpha$ de una distribución ji-cuadrada con p grados de libertad [8].

4.2.2. Prueba de Wald

Si existiera evidencia estadística suficiente en la prueba del cociente de verosimilitudes para rechazar H_0, aún es posible que alguna de las variables regresoras no tenga influencia sobre la variable respuesta, i.e. que alguno de los parámetros β_j, $j = 0, 1, \ldots, p$, sea igual a cero. La *prueba de Wald* permite determinar individualmente si cada variable X_j es estadísticamente significativa para el modelo, $j = 0, 1, \ldots, p$:

$$H_0\colon \beta_j = 0 \qquad \textit{versus} \qquad H_1\colon \beta_j \neq 0.$$

El estadístico de prueba es

$$W_j = \frac{\hat{\beta}_j}{\hat{v}\left[\hat{\beta}_j\right]},$$

para $j \in \{1, \ldots, p\}$, tal que bajo H_0, $W_j \sim N(0,1)$ aproximadamente. Se rechaza H_0 a un nivel de significancia α si $W_j > z_{(1-\alpha)}$, con $z_{(1-\alpha)}$ el cuantil $1-\alpha$ de una distribución normal estándar [8].

4.3. Bondad de ajuste

En resumen, los métodos de ajuste de un modelo permiten obtener los valores de β que maximizan la función de verosimilitud para después seleccionar entre ellos los parámetros que corresponden a variables significativas para el modelo. Cuando ya se han incluido en el modelo las variables cuyos efectos principales e interacciones son significativas para describir el comportamiento de la variable respuesta, es natural preguntarse ¿qué tan bien se ajusta el modelo a los conjuntos de covariables de la muestra?

Considere el vector de valores observados $y^T = (y_1, y_2, \ldots, y_n)$ y el vector de valores estimados de la respuesta $\hat{y}^T = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ con el modelo de regresión obtenido, entonces si $(y - \hat{y})$ es un vector con entradas cercanas a cero, se dice que el modelo se ajusta correctamente a los datos [8].

La *bondad de ajuste* de un modelo es una cantidad que expresa qué tan bien el modelo en cuestión responde a los conjuntos de covariables existentes.

4.3.1. Devianza y AIC

Dado que los conjuntos de covariables x_i se encuentran relacionados con los valores observados y_i de la variable respuesta en los datos disponibles, $i = 1, \ldots, n$, suponga que existe algún β^* tal que el modelo generado permite obtener las respuestas asociadas a todos los conjuntos de covariables observados, entonces la verosimilitud del modelo con β^* es igual a 1.

Una manera de estimar la varianza de la variable respuesta, es comparar el modelo ajustado con el modelo saturado,

$$D = -2\log\left[\frac{l(\hat{\beta})}{l(\beta^*)}\right] = -2\left[L(\hat{\beta})\right].$$

D se conoce como la *devianza*, y dado que $\hat{\beta}$ maximiza la función de verosimilitud, entonces $\hat{\beta}$ minimiza la devianza. Por lo tanto el criterio de selección es elegir el modelo ajustado con la menor devianza [8].

La devianza es una medida de bondad de ajuste que se encuentra inversamente relacionada con la probabilidad de observar las respuestas dados los datos muestrales con el modelo ajustado a través de la función de verosimilitud (y log-verosimilitud), de modo que un valor más alto en la verosimilitud garantiza devianzas más pequeñas y mejores ajustes. Valores positivos de la devianza ocurren porque la función de verosimilitud es el producto de las funciones de probabilidad para cada individuo, de modo que este producto se encuentra en el intervalo $[0, 1]$, y entre más pobres sean los ajustes esta cantidad será menor, por lo que al aplicar la función logaritmo, la imagen sobre este dominio cae en los números negativos y al multiplicar por -2 de nuevo se obtiene un número positivo.

La devianza también se utiliza para formar el *criterio de información de Akaike* (AIC), que a menudo es reportado por los paquetes estadísticos y es una herramienta útil para la selección de un modelo,

$$\text{AIC} = \text{devianza} + 2\times(\text{número de parámetros incluidos en el modelo ajustado}).$$

El AIC penaliza el uso de parámetros redundantes para explicar los datos observados, es por esto que el criterio de decisión es seleccionar el modelo que proporcione el menor AIC. Por lo tanto, el objetivo general es encontrar el modelo que mejor se ajuste a los datos, prefiriendo aquellos modelos que tengan menos variables, este principio se conoce como el principio de *parsimonia*.

El AIC se utiliza en situaciones donde se comparan modelos estadísticos que compiten entre sí. La selección de un modelo basado en el AIC tiene implicaciones importantes. Mientras que las pruebas de hipótesis intentan encontrar el modelo "correcto" o "verdadero", el AIC esencialmente ignora la noción de un modelo "verdadero" y se restringe a encontrar el modelo "mejor ajustado" [4].

4.3.2. Prueba ji-cuadrada de Pearson

La *prueba ji-cuadrada de Pearson* sirve para medir la bondad de ajuste de un cierto modelo propuesto, cuando se tiene un número acotado de conjuntos de covariables y el tamaño n de la población no es demasiado grande. Se utiliza también para estudios que no tienen variables regresoras continuas.

Sea J el número total de conjuntos de covariables en los datos, y sea m_j el número de personas que tienen el conjunto de covariables x_j, $j = 1, 2, \ldots, J$. Es fácil ver que $\sum_{i=1}^{J} m_j = n$.

Sea y_j el número de respuestas $y = 1$ entre los m_j sujetos, i.e. el número de personas con el mismo conjunto de covariables x_j y respuesta positiva.

Es posible estimar y_j como

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \, \frac{e^{\hat{\beta}^T x_j}}{1 + e^{\hat{\beta}^T x_j}}.$$

Para un conjunto particular x_j de covariables, el residuo de Pearson es

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j \, (1 - \hat{\pi}_j)}}. \tag{4.4}$$

La prueba de hipótesis es la siguiente:

H_0: El modelo está correctamente ajustado.

versus

H_1: El modelo no está correctamente ajustado.

El estadístico de prueba es la suma de los cuadrados de los residuos de (4.4),

$$\begin{aligned} X^2 &= \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2 \\ &= \sum_{j=1}^{J} \frac{(y_j - m_j \hat{\pi}_j)^2}{m_j \hat{\pi}_j \, (1 - \hat{\pi}_j)}, \end{aligned}$$

tal que $X^2 \sim {\chi^2}_{(n-(k+1))}$ bajo H_0, donde k es el número de parámetros del modelo ajustado. Se rechaza H_0 a un nivel de significancia α si $X^2 > {\chi^2}_{(n-(k+1),\ 1-\alpha)}$, con ${\chi^2}_{(n-(k+1),\ 1-\alpha)}$ el cuantil $1 - \alpha$ de una distribución ji-cuadrada con $n - k - 1$ grados de libertad [8].

4.3.3. Análisis de los residuos

Debido a la aleatoriedad que rige los datos observados, existe la posibilidad de que el modelo especificado no sea el más adecuado en cuanto a la relevancia de las variables (generalmente determinada por el especialista del área de aplicación), o la suficiencia de sus combinaciones lineales para explicar dichos datos. Es por esto que se "mide" la precisión del modelo por medio de los *residuos*.

4.3.3.1. Devianza residual

La *devianza residual* se define como

$$d(y_j, \hat{\pi}_j) \;=\; \pm\sqrt{2\left[y_j \log\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\log\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]}\,,$$

donde $\pm$ corresponde al signo de $(m_j - y_j)$.

En el caso en que no hay respuestas positivas en los individuos con el conjunto de covariables x_j, i.e. $y_j = 0$, la devianza residual se define como

$$d(y_j, \hat{\pi}_j) \;=\; -\sqrt{2m_j \mid \log(1-\hat{\pi}_j) \mid}\,.$$

En el caso en que todos los individuos cuyo conjunto de covariables sea x_j presenten respuestas positivas, i.e. $y_j = m_j$, la devianza residual se define como

$$d(y_j, \hat{\pi}_j) \;=\; \sqrt{2m_j \mid \log(\hat{\pi}_j) \mid}\,.$$

La prueba de hipótesis es la misma que en la prueba ji-cuadrada de Pearson.

El estadístico de prueba es la suma de los cuadrados de las devianzas residuales,

$$D_r \;=\; \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2,$$

tal que $D_r \sim {\chi^2}_{(J-(p+1))}$ bajo H_0. Cuando hay tantos conjuntos de covariables como individuos, i.e. $J = n$, $D_r = D$ la devianza definida en la sección 4.3.1. En tal caso, esta estadística no sirve para validar o refutar la hipótesis planteada.

Se rechaza H_0 a un nivel de significancia α si $D_r > {\chi^2}_{(J-(p+1),\ 1-\alpha)}$, con ${\chi^2}_{(J-(p+1),\ 1-\alpha)}$ el cuantil $1-\alpha$ de una distribución ji-cuadrada con $J-p-1$ grados de libertad [8].

4.3.4. Prueba de Hosmer-Lemeshow

A menudo la prueba ji-cuadrada de Pearson puede presentar problemas cuando el modelo ajustado contiene alguna variable continua, en cuyo caso el número de conjuntos de covariables distintos será $J = n$ o $J \approx n$, lo cual es peor cuando el número de individuos en observación es grande.

Por ello Hosmer y Lemeshow propusieron la *prueba de Hosmer-Lemeshow*, la cual estima las probabilidades $\hat{\pi}_j$, $j = 1, \ldots, J$, de cada uno de los conjuntos de covariables observados y las ordena de mayor a menor. Luego se agrupan las probabilidades estimadas en g grupos de igual tamaño, cada grupo contiene $n'_k = n/g$ individuos, $k = 1, \ldots, g$. Cada grupo formado constituye un cuantil.

La prueba de hipótesis es la siguiente:

H_0: El modelo está correctamente ajustado.

versus

H_1: El modelo no está correctamente ajustado.

El estadístico de prueba es

$$\widehat{C} \;=\; \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)},$$

donde n'_k es el número de individuos en el k-ésimo grupo,

$$o_k \;=\; \sum_{j=1}^{c_k} y_j,$$

donde c_k es el número de conjuntos de covariables en el k-ésimo cuantil y o_k es el número de respuestas $y = 1$ en los c_k conjuntos de covariables, y

$$\bar{\pi}_k \;=\; \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$$

es la probabilidad promedio estimada del k-ésimo cuantil.

Bajo H_0, $\widehat{C} \sim {\chi^2}_{(g-2)}$. Se rechaza H_0 a un nivel de significancia α si $\widehat{C} > {\chi^2}_{(g-2,\ 1-\alpha)}$, con ${\chi^2}_{(g-2,\ 1-\alpha)}$ el cuantil $1 - \alpha$ de una distribución ji-cuadrada con $g - 2$ grados de libertad [8].

Capítulo 5

Aplicación

5.1. Introducción

La *enfermedad de Parkinson* (*EP*) es el segundo desorden neurodegenerativo más común a nivel mundial, sólo después de la enfermedad de Alzheimer [13]. El *factor de riesgo* más impactante en la incidencia de la EP es la *edad* [18]. La EP rara vez es diagnosticada o aparece antes de los 50 años de edad y su incidencia se incrementa rápidamente después de los 60 años, siendo ligeramente mayor en hombres que en mujeres, sin embargo, la edad promedio en que la EP es diagnosticada es a los 70 años [18]. He aquí un primer acercamiento al problema, el lapso de tiempo en que en promedio aparece la enfermedad y su diagnóstico es de diez años, esto da oportunidad al avance neurodegenerativo de dicha enfermedad. Para una persona de 70 años, la edad puede representar en sí misma un impedimento para el diagnóstico de la EP, pues el traslado al hospital y los costos que esto le genera, incluyendo la consulta física, le resultan generalmente inconvenientes.

Una de las razones por las que el margen entre aparición y diagnóstico de la enfermedad es tan amplio, es que algunos de los síntomas de la EP normalmente no se presentan al mismo tiempo ni en etapas tempranas de la enfermedad o pueden ser confundidos con los síntomas de alguna otra patología, provocando que el inicio de un tratamiento específico sea aplazado. Los síntomas de la EP son mayormente motores en su naturaleza: temblor en reposo, rigidez muscular, *bradicinesia* o lentitud del movimiento, *acinesia* o incapacidad para iniciar un movimiento y alteraciones de la postura [7], trastornos generales del movimiento; y deterioro cognitivo [16]. Sin embargo, estos mismos síntomas se presentan en el *parkinsonismo*, que es causado por el uso de ciertos fármacos [7] y exposición a neurotoxinas [16].

La *disartria hipocinética* es uno de los primeros síntomas de la EP, la cual es un desorden del habla caracterizado por problemas con la articulación normal del habla y un cierto

retraso o torpeza al hablar. El 98 % de los casos de disartria es causado por la EP. El trastorno motor del habla es resultado de los efectos neuropatológicos de la EP sobre el sistema motor del habla provocando cambios en el control y la producción del habla, por lo que suele utilizarse la presencia de disartria en el fallo del diagnóstico médico del padecimiento neurológico. La disartria es comúnmente el primer síntoma de la EP [7]. La *disfonía* es otro trastorno asociado con la EP, el cual produce síntomas tales como reducción de la sonoridad, dificultad respiratoria, torpeza, disminución de la energía y exagerado temblor al hablar [9].

Dado que el 90 % de las personas con EP padece trastornos del habla [16], los cambios producidos por la enfermedad pueden ser medidos acústicamente para el diagnóstico de ésta, debido a que en tales cambios existen parámetros sensitivos del habla que son detectables hasta 5 años antes del diagnóstico de la enfermedad [7]. Lo anterior ha llevado a los investigadores de la EP a desarrollar herramientas auxiliares de diagnóstico, como modelos capaces de discriminar entre un individuo sano de uno con EP a través de los datos provenientes de grabaciones de voz en las que el individuo realiza fonaciones sostenidas con la vocal /a/ lo más estables que le sea posible durante al menos 5 segundos. El uso de fonaciones sostenidas en lugar de la articulación de palabras o frases previamente formadas se respalda en el argumento de que las métricas acústicas deben ser mínimamente afectadas por la variabilidad que se encuentra relacionada con el hablante y máximamente afectadas por el deterioro articulatorio provocado por la enfermedad [14].

5.2. Materiales y métodos

5.2.1. Los participantes

Un total de 80 sujetos mayores de 50 años de edad se contemplaron en el estudio. 40 de ellos sanos: 22 hombres (55 %) y 18 mujeres (45 %); y 40 de ellos con EP: 27 hombres (67.5 %) y 13 mujeres (32.5 %). La media ($\pm$ la desviación estándar) de la edad fue 66.38 ± 8.38 años para el grupo control (de individuos sin la enfermedad de Parkinson) y 69.58 ± 7.82 para el grupo de personas con EP. Los pacientes con EP presentaban al menos dos de los siguientes síntomas: temblor al reposo, bradicinesia o rigidez muscular [12]. Las personas con EP que participaron en el estudio son miembros de la Asociación Regional para la Enfermedad de Parkinson en Extremadura (España) [12].

5.2.2. Las grabaciones

Los individuos debían realizar una fonación sostenida a un volumen y tono cómodos, pronunciando la vocal /a/ de la manera más estable que les fuese posible. La fonación debía mantenerse por al menos 5 segundos y con una sola respiración. Cada individuo repetía este procedimiento tres veces y cada una de estas grabaciones fueron consideradas como repeticiones. Los datos fueron grabados utilizando una computadora portátil con una tarjeta de sonido externa (TASCAMUS322) y un micrófono de diadema (AKG520). La grabación digital fue hecha a una velocidad muestral de 44.1KHz y a una resolución de 16bits/muestra utilizando el software Audacity (versión 2.0.5) [12].

5.2.3. Extracción de las características

La base de datos se compone de 44 variables o características acústicas, que pueden ser clasificadas en cinco grupos:

1. Jitter (medidas de perturbación del tono local [12] o frecuencia [17]).
2. Shimmer (medidas de perturbación de la amplitud [12], [17]).
3. Características del ruido [12] (proporción de señales a ruido [17]).
4. Mediciones de las frecuencias cepstrales de Mel [17].
5. Mediciones no lineales [12].

La composición de algunos de los grupos anteriores se muestra en los siguientes cuadros.

Cuadro 5.1: Medidas de perturbación del tono local [12], [17].

Variable	Descripción
Jitter_rel	Perturbación relativa del tono local (expresada en porcentaje).
Jitter_abs	Perturbación absoluta del tono local.
Jitter_RAP	Perturbación relativa promedio del tono local.
Jitter_PPQ	Cociente de perturbación del tono local.

Cuadro 5.2: Medidas de perturbación de la amplitud [12], [17].

Variable	Descripción
Shimmer_loc	Perturbación local de la amplitud.
Shimmer_dB	Perturbación de la amplitud en decibeles.
Shimmer_APQ3	Cociente de perturbación de 3 puntos de amplitud.
Shimmer_APQ5	Cociente de perturbación de 5 puntos de amplitud.
Shimmer_APQ11	Cociente de perturbación de 11 puntos de amplitud.

La proporción de armonías en los ruidos (HNR, por sus siglas en inglés) es una medición del relativo nivel de ruido presente en el habla. Existen varios tipos de HNR constituidos en base a su enfoque, a saber, aquellos cuyo dominio es el tiempo o su dominio es la frecuencia. En esta muestra, se han obtenido según diferentes anchuras de intervalos de frecuencia [12].

Cuadro 5.3: Características del ruido [12].

Variable	Descripción
HNR05	Proporción de armonías a ruidos entre 0 y 500 Hz.
HNR15	Proporción de armonías a ruidos entre 0 y 1500 Hz.
HNR25	Proporción de armonías a ruidos entre 0 y 2500 Hz.
HNR35	Proporción de armonías a ruidos entre 0 y 3500 Hz.
HNR38	Proporción de armonías a ruidos entre 0 y 3800 Hz.

Los coeficientes cepstrales de frecuencias de Mel (MFCC, por sus siglas en inglés) se encuentran relacionados con el espectro del habla y dependen de la posición de los miembros articuladores [12]. Dichos coeficientes detectan pequeños cambios en la posición de los miembros articuladores (la lengua y los labios) debidos al temblor provocado por la EP cuando se realiza una fonación sostenida [17], [12]. Para cada uno de los participantes se consideraron 13 MFCC's (MFCC0, MFCC1, ... , MFCC12). Se registraron también características "Delta" (Delta0, Delta1, ... , Delta12) derivadas del tiempo de los MFCC's, de modo que cada grabación tiene 26 mediciones de este tipo (13 MFCC's y 13 coeficientes Delta) asociadas, que son utilizadas para caracterizar la EP [12].

La presencia de mediciones de características acústicas no lineales es de particular importancia porque las voces patológicamente con disfonía severa son aquellas que más comúnmente presentan fenómenos aleatorios no lineales, mientras que las voces de individuos sanos se encuentran más relacionadas con el modelo lineal [12].

Cuadro 5.4: Mediciones no lineales [12].

Variable	Descripción
RPDE	Entropía de la densidad del periodo de recurrencia.
DFA	Análisis de fluctuación sin tendencias.
PPE	Entropía del periodo del tono.
GNE	Cociente de excitación glotal sobre el ruido.

La entropía de la densidad del periodo de recurrencia (RPDE, por sus siglas en inglés) estima la duración de los ciclos de las cuerdas vocales en el tiempo transcurrido entre colisiones de cuerdas vocales contiguas [17]. La entropía del periodo del tono (PPE) es una log-transformación de la frecuencia fundamental [16] que sirve como una medición de la disfonía [9]. El cociente de excitación glotal sobre el ruido (GNE) pretende cuantificar la cantidad de excitación vocal causada por la oscilación de las cuerdas vocales contra la excitación producida por el ruido de turbulencia [12].

Cada individuo posee 3 conjuntos de covariables que llamaremos *repeticiones* por tratarse de conjuntos de covariables asociados al mismo individuo, de modo que se obtienen 240 registros vocales con 44 atributos o características acústicas cada uno.

5.3. Metodología

La base de datos cuenta con 44 covariables acústicas medibles como se expuso anteriormente, sin embargo, dado que el número de covariables es "grande" y existe correlación entre algunas de ellas, especialmente entre aquellas pertenecientes a un mismo grupo, excepto en el grupo de mediciones no lineales, se reducirá la dimensión por medio del análisis de componentes principales comunes sobre cada grupo y posteriormente se ajustará un modelo de regresión logística como herramienta probabilística para discriminar entre un individuo sano y uno con EP, para su diagnóstico.

Con el fin de identificar las variables que estadísticamente denotan a una persona como sana o con EP, se obtienen los CPC de las matrices de covarianzas de las repeticiones en cada uno de los siguientes grupos:

1. Grupo Jitter (4 variables): Jitter_rel, Jitter_abs, Jitter_RAP y Jitter_PPQ.
2. Grupo Shimmer (5 variables): Shimmer_loc, Shimmer_dB, Shimmer_APQ3, Shimmer_APQ5 y Shimmer_APQ11.
3. Grupo HNR (5 variables): HNR05, HNR15, HNR25, HNR35 y HNR38.

4. Grupo MFCC (13 variables): MFCC0, MFCC1, ..., MFCC12.

5. Grupo Delta (13 variables): Delta0, Delta1, ..., Delta12.

Se eligirán los CPC que conserven la mayor variabilidad de los datos de cada grupo y sean capaces de explicar el modelo en al menos un 90 %. A las componentes obtenidas las llamaremos *índices*.

Es posible pensar que la dimensión de cada uno de los grupos se reducirá mediante el ACPC, ya que las correlaciones entre las variables dentro de los grupos son altas.

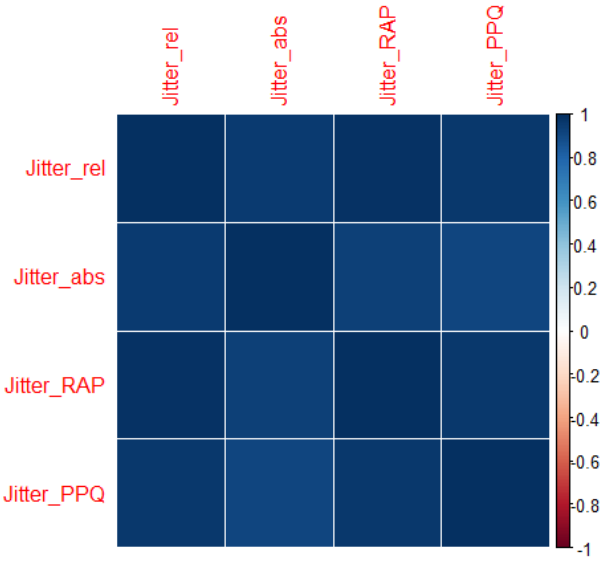


FIGURA 5.1: Gráfica de correlaciones en el grupo Jitter.

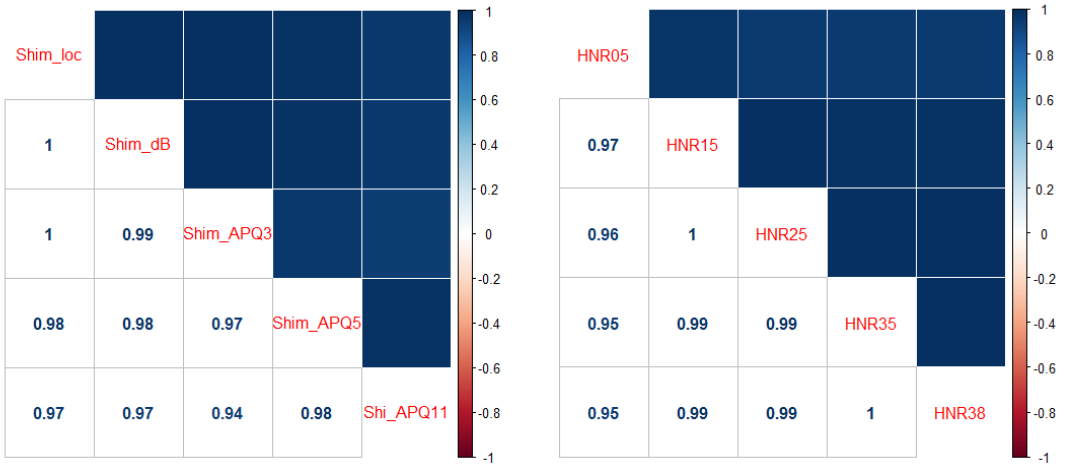


FIGURA 5.2: Gráfica de correlaciones en el grupo Shimmer (a la izquierda) y en el grupo HNR (a la derecha).

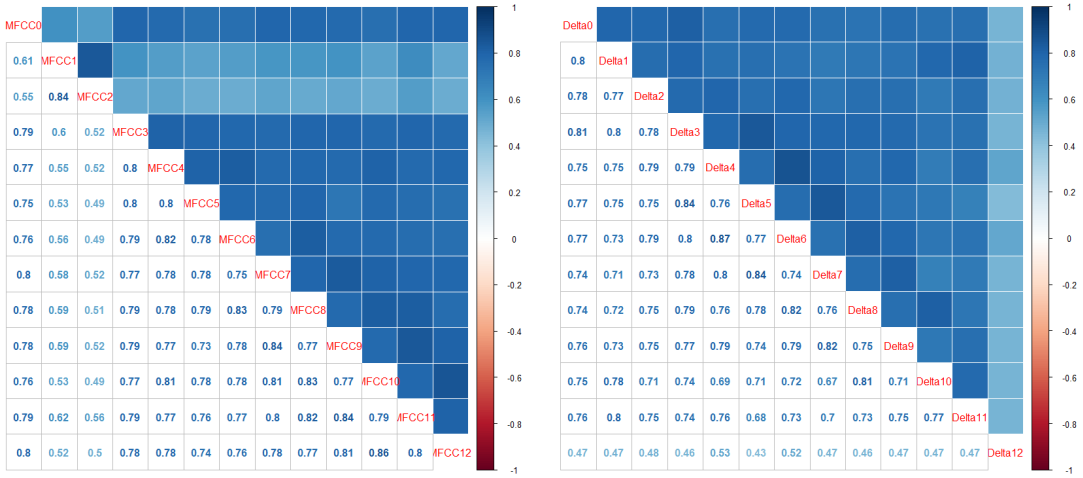


FIGURA 5.3: Gráfica de correlaciones en el grupo MFCC (a la izquierda) y en el grupo Delta (a la derecha).

Según es observable, las correlaciones son más altas en los grupos Jitter, Shimmer y HNR. Esto tiene un impacto directo con el número de componentes que se seleccionan de un grupo, i.e. que entre más correlacionadas se encuentran las variables, la variabilidad de los datos se explica con una cantidad menor de índices, afectando de manera directa, y en cierto sentido podría decirse que de manera "positiva", la dimensión de la muestra después del ACPC.

Dentro del grupo de mediciones no lineales podemos considerar las variables como "aisladas", de modo que en el modelo de regresión logística se incluirán todas las mediciones no lineales junto con los índices obtenidos, a fin de construir *biomarcadores* que permitan diagnosticar correctamente a personas con EP mediante señales acústicas.

5.3.1. CPC de la muestra

Debido a que la muestra consta de 3 grabaciones por cada persona incluida en el estudio, las cuales llamamos repeticiones, se dividió la base de datos en tres grupos, a saber: primera, segunda y tercera repetición. Dentro de cada submuestra de repeticiones se identifican los 5 grupos previamente mencionados (Jitter, Shimmer, HNR, MFCC y Delta) y obtenemos las matrices de covarianzas para cada uno de los grupos en las tres submuestras.

Denotamos como $\mathbf{\Psi}_i^A$ a la matriz de covarianzas del grupo "A", A={Jitter, Shimmer, HNR, MFCC y Delta}, perteneciente a la repetición "i", $i = \{1, 2, 3\}$. Entonces obtenemos los componentes principales comunes para cada grupo. Precisamente la razón de realizar el ACPC, y no simplemente el análisis de componentes principales, es que

deseamos obtener los componentes principales de cada grupo que sean comunes a cada una de las tres repeticiones.

Con la intención de clarificar el procedimiento seguido, para el grupo Jitter, por ejemplo, las matrices cuadradas de dimensión 4 (porque el grupo Jitter consta de cuatro mediciones) son $\mathbf{\Psi}_1^{Jitter}$, $\mathbf{\Psi}_2^{Jitter}$ y $\mathbf{\Psi}_3^{Jitter}$. Conforme a la teoría del ACPC, para estas matrices obtenemos una misma base ortogonal $\beta \in \mathbb{R}^4$, cuyas columnas son los eigenvectores asociados a cada una de las componentes del grupo en consideración como en (3.4), y tres matrices diagonales distintas como en (3.5), una por cada submuestra de repeticiones, cuyos elementos son los eigenvalores asociados a las componentes del grupo que se está analizando, en este ejemplo, Jitter.

Denotamos como $\lambda_{i,j}^A$ al j-ésimo eigenvalor del grupo A y submuestra de repeticiones i. Recordando también que $\lambda_{i,1}^A \geq \lambda_{i,2}^A \geq \ldots \geq \lambda_{i,L}^A$, para un grupo con L componentes y repeticiones $i = \{1, 2, 3\}$. En este trabajo deseamos incluir solamente las componentes que nos permitan conservar al menos el 90 % de la variabilidad existente entre las componentes del grupo cuando interactúan entre sí. Además, es posible pensar que se reducirá la dimensión de los grupos, toda vez que encontramos que las correlaciones entre las variables de cada grupo son altas.

Para conocer la variabilidad que las componentes explican o aportan al grupo, realizamos el cociente

$$\frac{\sum_{j=1}^{k} \lambda_{i,j}^A}{\sum_{j=1}^{L} \lambda_{i,j}^A} .$$

Si deseamos conocer la variabilidad explicada por los primeros k componentes, realizamos la suma parcial de los primeros k eigenvalores en el numerador y, bajo cualquier situación, la suma del denominador se realiza sobre todos los eigenvalores, $j = \{1, 2, \ldots, L\}$, para un grupo con L componentes.

Para el grupo Jitter, los valores propios obtenidos relacionados con las cuatro componentes del grupo son:

	Repetición 1	Repetición 2	Repetición 3
[1]	1.484221e-01	6.099548e-01	1.027066e-01
[2]	2.597852e-07	5.446505e-06	4.396545e-07
[3]	1.310743e-07	3.162110e-07	2.447808e-07
[4]	1.574788e-10	1.450113e-10	1.058978e-10

En nuestro ejemplo se tienen los siguientes datos de las sumas parciales de este cociente para el grupo Jitter y para las repeticiones 1, 2 y 3, respectivamente:

```
[1] 0.9999974  0.9999991  1.0000000  1.0000000
[2] 0.9999906  0.9999995  1.0000000  1.0000000
[3] 0.9999933  0.9999976  1.0000000  1.0000000
```

Es decir, que para el grupo Jitter, la primer componente expresa más del 99.99 % de la variabilidad presente en el grupo en la primera, segunda y tercera repetición, de modo que podemos seleccionar sólo la primer componente como la componente principal común a todas las repeticiones, despreciando las demás componentes del grupo. Una vez que hemos escogido esta componente, obtenemos la combinación lineal que formará el índice para el grupo Jitter. Para lograr esto, multiplicamos la matriz del grupo Jitter por el primer vector de la base ortogonal β obtenida por el ACPC, puesto que este primer vector columna corresponde a la primer componente. Hacemos esto para cada una de las repeticiones de modo que obtenemos los valores del CPC del grupo Jitter en cada repetición. No obstante que existen métodos con los cuales es posible determinar qué componente(s) de las repeticiones utilizar en cada uno de los grupos, creemos que tales herramientas matemáticas exceden los propósitos de este trabajo y por esta razón hemos escogido únicamente obtener el promedio de los tres componentes para cada individuo. Nuestro primer CPC es el índice CPC-Jitter.

Realizando un procedimiento similar, en el grupo Shimmer, la primer componente expresa el 99.91 % de la variabilidad del grupo, por ello la seleccionamos como la componente principal común y obtenemos los valores de su combinación lineal para expresar los datos del grupo Shimmer. Al índice obtenido lo llamamos CPC-Shimmer.

Análogamente seleccionamos la primer componente del grupo HNR, pues explica el 98.49 % de la variabilidad presente en el grupo HNR. CPC-HNR es el tercer índice.

En el grupo MFCC no es suficiente conservar una sola componente, sino seis para expresar la variabilidad que deseamos preservar en los datos (mayor o igual al 90 %), sin embargo, en aras del principio de parsimonia incluimos solamente las primeras dos componentes, que expresan poco más del 80 % de la variabilidad del grupo. Los siguientes datos son la variabilidad aportada por los primeros eigenvalores del grupo para las repeticiones 1, 2 y 3, respectivamente:

```
[1] 0.6856267  0.8113837  0.8399725  0.8637494  0.8842013  ...
[2] 0.7410061  0.8442991  0.8694891  0.8872070  0.9064866  ...
[3] 0.7658863  0.8382686  0.8624174  0.8842353  0.9015672  ...
```

Para obtener las combinaciones lineales de los primeros dos componentes, multiplicamos la matriz del grupo MFCC por el primer y segundo vector de la base ortogonal β. A los índices obtenidos los llamamos CPC1-MFCC y CPC2-MFCC.

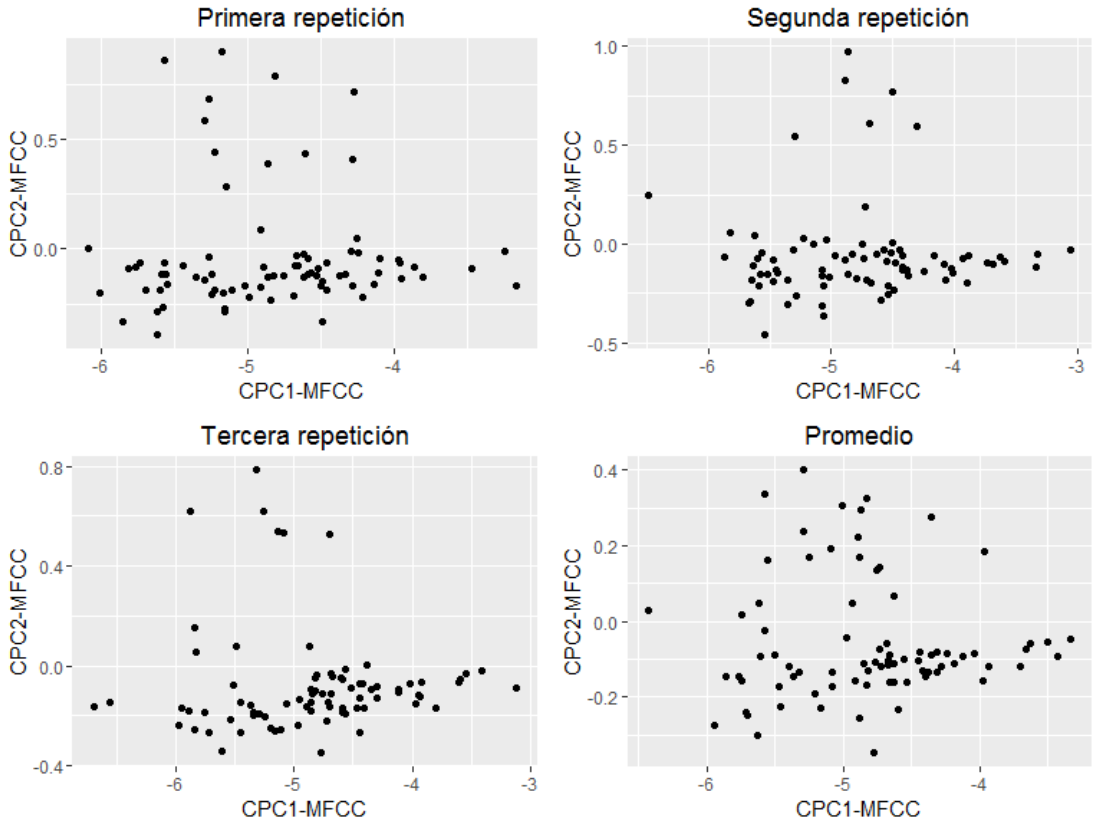


FIGURA 5.4: Gráficas de los componentes principales comunes del grupo MFCC. (Dimensión inicial del grupo: 13; dimensión obtenida tras el ACPC: 2.)

Con el grupo Delta sucede algo parecido, sin embargo, se escogen únicamente los primeros dos componentes a pesar de aportar sólo un poco más que el 78 % de la variabilidad. A los índices obtenidos los llamamos CPC1-Delta y CPC2-Delta, en ese mismo orden.

La variabilidad aportada por las primeras componentes, para la primera, segunda y tercera repetición, respectivamente, es:

```
[1] 0.7412348  0.7933562  0.8186717  0.8511582  0.8791112  ...
[2] 0.7251518  0.7801935  0.8310371  0.8560636  0.8865226  ...
[3] 0.7611091  0.8293353  0.8591836  0.8852024  0.9038717  ...
```

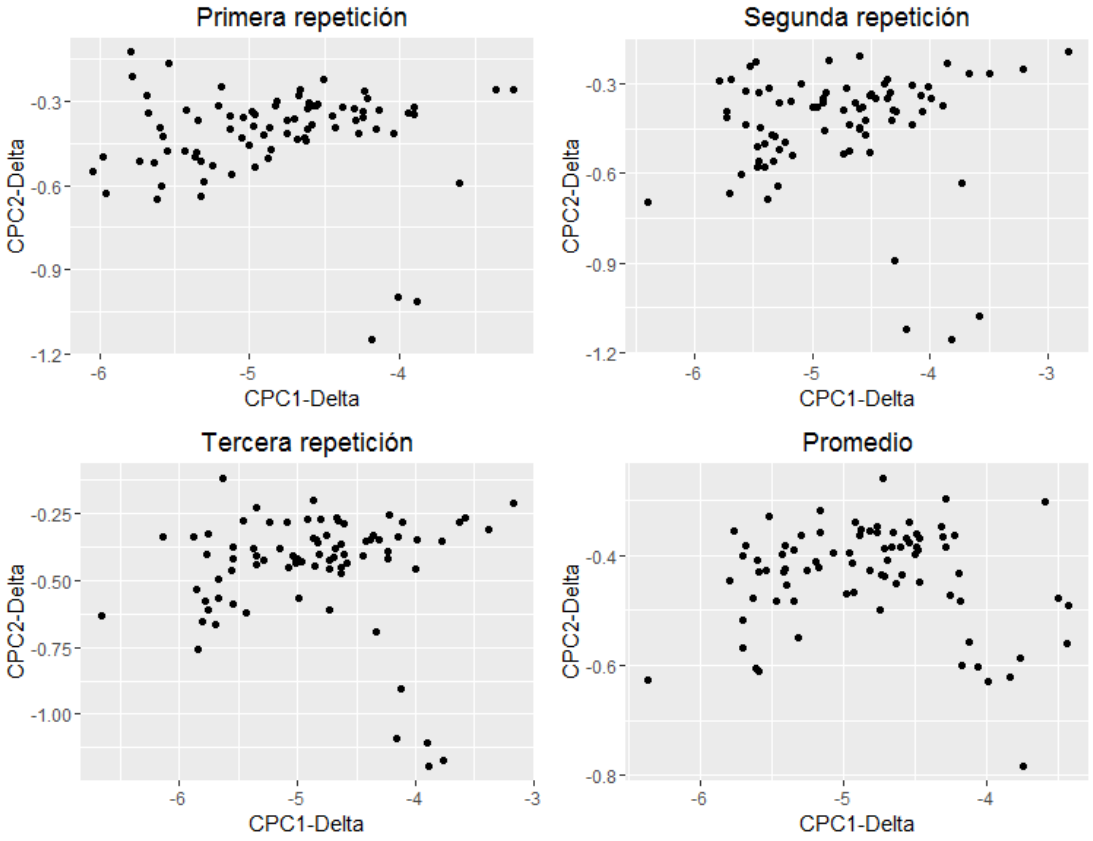


FIGURA 5.5: Gráficas de los componentes principales comunes del grupo Delta. (Dimensión inicial del grupo: 13; dimensión obtenida tras el ACPC: 2.)

Incorporando las mediciones no lineales de las variables aisladas, la muestra de biomarcadores contiene los siguientes índices: Sexo, CPC-Jitter, CPC-Shimmer, CPC-HNR, RPDE, DFA, PPE, GNE, CPC1-MFCC, CPC2-MFCC, CPC1-Delta y CPC2-Delta.

5.3.2. Regresión logística

Ahora que se ha reducido la dimensión y se han seleccionado variables e índices como biomarcadores a través del ACPC, podemos estar seguros que esta submuestra contiene las variables suficientes para hacer "la mejor reconstrucción" de la muestra original en términos de la variabilidad presentada entre los datos que teníamos al principio.

Es posible formar un modelo lineal generalizado asociado a la variable respuesta con las variables presentes en el biomarcador. La respuesta en que estamos interesados es, desde luego, el estado del individuo, i.e. si el participante posee o no la EP. Es natural pensar en tal planteamiento debido a que deseamos conocer el grado en que el habla y ciertas mediciones acústicas del habla pueden ser un método seguro de diagnóstico de la EP. Dado que la variable "Estado" es una variable dicotómica, i.e. toma sólo dos valores, "0" cuando el individuo no tiene la EP y "1" cuando sí, cada individuo tiene

asociada una función de masa de probabilidad Bernoulli de parámetro $\pi(\mathbf{X})$, hasta este momento desconocido. Por otro lado, la liga canónica asociada a una variable aleatoria Bernoulli es, como hemos visto, la liga conocida como *logit*. Con base en esta justificación, consideramos el modelo de regresión logística como el modelo a partir del cual nos será posible discriminar, según su conjunto de covariables, entre un participante con EP y uno sin la enfermedad.

A continuación se presentan los coeficientes de regresión.

(Intercepto)	Sexo	CPC-Jitter	CPC-Shimmer	CPC-HNR
67.93521	-5.31299	21.52507	-1.56493	0.08686
RPDE	DFA	PPE	GNE	CPC1-MFCC
-79.74077	-24.88976	0.49656	8.64356	6.64838
CPC2-MFCC	CPC1-Delta	CPC2-Delta		
-14.74723	0.29137	-3.88439		

Por lo tanto, el modelo propuesto es el siguiente:

$$\begin{aligned} logit(\pi(\mathbf{X})) &= 67.9352 - 5.3129 \cdot \mathbf{Sexo} + 21.5250 \cdot \mathbf{CPC\text{-}Jitter} \\ &- 1.5649 \cdot \mathbf{CPC\text{-}Shimmer} + 0.0868 \cdot \mathbf{CPC\text{-}HNR} \\ &- 79.7407 \cdot \mathbf{RPDE} - 24.8897 \cdot \mathbf{DFA} + 0.4965 \cdot \mathbf{PPE} \\ &+ 8.6435 \cdot \mathbf{GNE} + 6.6483 \cdot \mathbf{CPC1\text{-}MFCC} \\ &- 14.7472 \cdot \mathbf{CPC2\text{-}MFCC} + 0.2913 \cdot \mathbf{CPC1\text{-}Delta} \\ &- 3.8843 \cdot \mathbf{CPC2\text{-}Delta} \;=\; \beta^T \mathbf{X} \end{aligned}$$

Lo anterior significa que, fijando el valor del resto de las variables cuando no están siendo consideradas, un cambio unitario en la variable: CPC-Jitter incrementa las log-posibilidades[1] (de tener la EP a no tenerla) en 21.52 unidades; CPC-Shimmer las reduce en 1.56 unidades; CPC-HNR las incrementan en 0.08 unidades, mientras que RPDE y DFA las disminuyen en 79.74 y 24.88 unidades, respectivamente; PPE, GNE y CPC1-MFCC las incrementan en 0.49, 8.64 y 6.64 unidades, respectivamente; CPC2-MFCC las reduce en 14.74 unidades, mientras CPC1-Delta las incrementa en 0.29 unidades y CPC2-Delta las reduce en 3.88 unidades. En el caso de la variable Sexo, al fijar el valor de las demás variables, las log-posibilidades se reducen en 5.31 unidades si se trata de mujeres, en comparación con las log-posibilidades en el caso de los hombres en que la variable "Sexo" toma el valor cero. Ahora es posible obtener la probabilidad $\pi(\mathbf{X})$ como en (4.1). En el capítulo 4 se expusieron algunas de las propiedades de este modelo matemático, las cuales analizaremos a continuación.

[1] $logit\left(\pi(\mathbf{X})\right) = \log\left[\pi(\mathbf{X})/\left(1-\pi(\mathbf{X})\right)\right]$.

5.3.3. Ajuste del modelo

El ACPC ayudó a reducir la dimensión a 7 índices dentro de un total de 40, conservando la mayor variabilidad total de los datos y la mayor información relevante, o de otro modo, eliminando información redundante, sin embargo, luego introdujimos las variables de mediciones no lineales, las cuales no estaban altamente relacionadas entre sí, de manera que una vez que se han ajustado los parámetros del modelo, todavía es importante identificar cuáles de ellos son estadísticamente significativos (marginalmente).

Para conocer cuáles de los parámetros estimados son estadísticamente significativos al realizar la prueba de Wald, se obtienen los intervalos de confianza, para lo que se requiere el error estándar estimado de cada parámetro ($\hat{v}[\hat{\beta}_j]$).

La estimación del intervalo del 95 % de confianza para β_j es $\hat{\beta}_j \pm z_{0.975} \cdot \hat{v}[\hat{\beta}_j]$, donde $z_{0.975}$ es el cuantil 0.975 de una distribución normal estándar.

Coeficientes:

	Estimado	Error est.	2.5 %	97.5 %	Pr(>|z|)	
(Intercepto)	67.93521	37.09970	6.001916	154.639156	0.06708	.
Sexo	-5.31299	2.21038	-10.647797	-1.727762	0.01623	*
CPC-Jitter	21.52507	9.45479	6.371989	45.046913	0.02281	*
CPC-Shimmer	-1.56493	8.31698	-22.049669	13.851980	0.85075	
CPC-HNR	0.08686	0.04271	0.015304	0.189696	0.04197	*
RPDE	-79.74077	30.60981	-160.069633	-32.781086	0.00919	**
DFA	-24.88976	15.02083	-59.176467	0.973815	0.09752	.
PPE	0.49656	3.36832	-6.128723	7.860302	0.88280	
GNE	8.64356	30.41854	-48.886711	75.305729	0.77629	
CPC1-MFCC	6.64838	6.65077	-5.582173	21.936135	0.31748	
CPC2-MFCC	-14.74723	6.60902	-30.889835	-4.257386	0.02566	*
CPC1-Delta	0.29137	6.14222	-12.924290	12.412968	0.96216	
CPC2-Delta	-3.88439	8.80909	-22.922247	13.668203	0.65925	

Desde aquí es fácil observar qué variables o índices resultan ser significativas mientras se encuentren las demás variables en el modelo: aquellas cuyo p-value es menor que la región de rechazo, en este caso, las variables Sexo, CPC-Jitter, CPC-HNR, RPDE y CPC2-MFCC.

5.3.3.1. Prueba del cociente de verosimilitudes

La comparación entre el modelo ajustado y un modelo nulo proporciona una medida de la importancia que tienen las variables incluidas en el modelo para explicar la respuesta, con respecto a un modelo en que no se considera la inclusión de ninguna de las variables. La hipótesis en prueba es que ninguna de las variables es estadísticamente significativa ($H_0 : \beta = 0$) contra que al menos una lo es ($H_1 : \beta \neq 0$).

De modo que G = – $2(L(\beta) - L(\hat{\beta})) = 110.904 - 34.134 = 76.769$, y $\chi^2_{(13;0.95)} = 22.36203$. Como G> $\chi^2_{(13,0.95)}$, se rechaza la hipótesis nula H_0, es decir que no existe evidencia estadística suficiente para afirmar que en general, las variables en el modelo son estadísticamente iguales a cero.

5.3.3.2. Prueba de Wald

Lo anterior no significa que no exista algún parámetro estadísticamente igual a cero en el modelo. La hipótesis que ahora se prueba es que alguno de los coeficientes (β_j, $j = 1, \ldots, 13$) no sea estadísticamente significativo ($H_0 : \beta_j = 0$) contra que sí lo sea ($H_1 : \beta_j \neq 0$).

Prueba de Wald

	Ji-cuadrada	gl	P(> X2)
(Intercepto)	3.4	1	0.067
Sexo	5.8	1	0.016
CPC-Jitter	5.2	1	0.023
CPC-Shimmer	0.035	1	0.85
CPC-HNR	4.1	1	0.042
RPDE	6.8	1	0.0092
DFA	2.7	1	0.098
PPE	0.022	1	0.88
GNE	0.081	1	0.78
CPC1-MFCC	1.0	1	0.32
CPC2-MFCC	5.0	1	0.026
CPC1-Delta	0.0023	1	0.096
CPC2-Delta	0.19	1	0.66

De acuerdo con los resultados anteriores, la prueba de Wald nos asegura que no es posible rechazar la hipótesis nula en los casos en que el p-value (`P(>X2)`) es mayor que la

significancia de las pruebas (5 %). En el caso de las variables Sexo, CPC-Jitter, CPC-HNR, RPDE y CPC2-MFCC, la evidencia estadística es suficiente para rechazar que son variables no significativas para el modelo completo, mientras que para el resto de las variables no se rechaza la hipótesis nula, y las variables resultan ser no significativas. Es importante señalar que una variable puede no ser estadísticamente significativa en presencia de las demás variables del modelo, sin embargo, si eliminamos alguna otra variable, la interacción se modifica, pudiendo ocurrir que la variable que no era considerada como significativa, ahora lo sea.

5.3.4. Bondad de ajuste

Una vez que se conocen las variables que conforman el modelo, es natural preguntarse ¿qué tan bien se ajusta el modelo a los datos? Intuitivamente, responder a esta pregunta depende de la diferencia entre la respuesta proporcionada por el modelo ($\hat{y}$) y la respuesta observada en los datos (y), i.e. depende de los valores del vector ($y - \hat{y}$).

No es posible observar la utilidad de la devianza en este caso, pues es una herramienta de selección de modelos, cuando es posible comparar con otras devianzas. En tales casos se elige el modelo con menor devianza.

```
Devianza
34.13411
```

De igual manera, el AIC carece de interpretación en este caso, puesto que para aprovechar su utilidad como herramienta de selección de un modelo, es necesario compararlo con el AIC de otros modelos ajustados. Se elige el modelo con menor AIC.

```
AIC: 60.134
```

5.3.5. Análisis de los residuos

5.3.5.1. Devianza residual

La hipótesis nula de esta prueba es la misma que en la prueba de la ji-cuadrada de Pearson. Como se explicó en la sección correspondiente, cuando hay tantos conjuntos de covariables distintos como individuos, la devianza residual es la misma que la devianza del modelo obtenida por log-verosimilitud. El estadístico de prueba es la suma de los cuadrados de las devianzas residuales.

```
Devianza residual:   42.735
```

```
Cuantil 95 de una distribución ji-cuadrada con 66 grados de libertad:
85.96491
```

Como $42.735 = D_r < \chi^2_{(80-12-1,0.95)} = 85.96491$, no se rechaza la hipótesis nula. Es decir que no existe evidencia suficiente para rechazar que el modelo representa un buen ajuste a los datos.

5.3.5.2. Prueba de Hosmer-Lemeshow

Esta prueba estima las probabilidades de que la variable respuesta tome el valor 1 para cada conjunto de covariables y las reúne en grupos de igual tamaño.

```
Pueba de bondad de ajuste de Hosmer y Lemeshow:
```

```
Ji-cuadrada = 0.35757, gl = 8, p-value = 1
```

El modelo ajusta bien los datos pues el p-value (1.0) es mayor al nivel de significancia de la prueba (5 %), de modo que el estadístico de prueba no cae en el área de rechazo. De lo anterior se puede inferir que, conforme al procedimiento estándar, se crean 10 grupos de probabilidades, por ello se tienen ocho grados de libertad y $\chi^2_{(8,0.95)} > \widehat{C}= 0.35757$. Por lo tanto, no existe evidencia estadística suficiente para rechazar que el modelo está bien ajustado.

5.3.6. Tabla de contingencia

Luego de haber hecho pruebas al modelo, es muy importante conocer lo efectiva que es realmente la prueba para diagnosticar la EP, para ello utilizamos una tabla de contingencia en donde veremos 4 posibles categorías dentro de las cuales puede caer el resultado de la prueba, esto nos ayudará a conocer si es un método factible para el diagnóstico de la EP en nuestra población.

Con los parámetros del modelo ajustado es posible conocer, conforme a su conjunto de covariables, la probabilidad de que el individuo obtenga un resultado negativo en la prueba (obtenga "0") o uno positivo (obtenga "1"). Pero, dado que $\pi(\mathbf{X})$ resulta ser cualquier número entre cero y uno, se asigna uno si $\pi(X) > 0.5$, y cero en caso contrario. Con base en estos resultados elaboramos la siguiente tabla de contingencia.

CUADRO 5.5: Tabla de contingencia con estructura de frecuencias.

	Diagnóstico		
Parkinson	Positivo	Negativo	Total
Sí	36	4	40
No	4	36	40
Total	40	40	80

5.3.6.1. Sensibilidad y especificidad

La sensibilidad de la prueba es la probabilidad de haber obtenido un diagnóstico positivo dado que se tiene la enfermedad, es decir, la proporción de enfermos correctamente identificados, en este caso

$$Sensibilidad = \frac{36}{36 \ + \ 4} = \frac{9}{10} = 90\,\%.$$

Por lo tanto podemos afirmar que el modelo ofrece un diagnóstico correcto en el 90 % de los casos cuando la persona tiene EP.

La especificidad de la prueba es la probabilidad de haber obtenido un diagnóstico negativo dado que no se tiene la enfermedad, es decir, la proporción de sanos correctamente identificados, y en este caso

$$Especificidad = \frac{36}{36 \ + \ 4} = \frac{9}{10} = 90\,\%.$$

Por lo tanto podemos afirmar que el modelo ofrece un diagnóstico correcto en el 90 % de los casos cuando la persona no tiene EP.

5.3.6.2. Riesgo relativo

Son las posibilidades de obtener un diagnóstico "Positivo" dado que se tiene la EP, sobre obtener el mismo diagnóstico sin tener la enfermedad. En este caso,

$$RR = \frac{\frac{36}{36 \ + \ 4}}{\frac{4}{36 \ + \ 4}} = \frac{36}{4} = 9,$$

es decir, que es 9 veces más probable obtener un resultado "Positivo" si se tiene la EP que si no se tiene. Un valor alto como éste se encuentra relacionado con el hecho de que la sensibilidad de la prueba es muy alta. Este resultado es favorable e indica que el método de diagnóstico utilizado es eficaz.

5.3.6.3. Razón de momios

Ya que es posible expresar la razón de momios en términos del riesgo relativo, tenemos que

$$RM = RR \times \frac{1-\pi_2}{1-\pi_1} = 9 \times \frac{1-\frac{4}{36+4}}{1-\frac{36}{36+4}} = 9 \times 9 = 81,$$

de esta manera podemos afirmar que con este método de diagnóstico es más probable obtener un resultado "Positivo" que uno "Negativo" teniendo la EP, que encontrarse en el mismo caso sin tener la enfermedad. Por lo tanto, podemos decir que el método de diagnóstico es eficiente.

Pero, para asegurar que existe una diferencia real entre obtener un resultado "Positivo" que uno "Negativo", es decir que $RM \neq 1$, obtenemos el intervalo de Wald de la RM. Si el intervalo no incluye al uno, aseguramos que la diferencia entre obtener un resultado u otro existe. La varianza del $\log(RM)$ es

$$\mathbb{V}\left(\log\left(\widehat{RM}\right)\right) \approx \left(\frac{1}{36}+\frac{1}{4}+\frac{1}{4}+\frac{1}{36}\right) = \frac{10}{18}.$$

El intervalo de Wald con el 95 % de confianza para $\log\left(\widehat{RM}\right)$ es

$$\log(81) \pm 1.959964\sqrt{\frac{10}{18}}$$

donde 1.959964 es el cuantil 0.975 de una distribución normal estándar.

Por lo tanto el intervalo con el 95 % de confianza para la RM es $(18.79476, 349.0866)$, y como el intervalo no incluye al uno, decimos que para el método de diagnóstico de la EP por medio de grabaciones de voz existe diferencia entre obtener un resultado "Positivo" sobre un "Negativo" cuando se tiene la EP que cuando no se tiene.

5.4. Resultados

La muestra original del estudio consta de 240 patrones de covariables, correspondientes a 80 personas mayores de 50 años (49 hombres y 31 mujeres) de entre las cuales 40 tienen EP y las otras 40 no presentan esta patología. Cada individuo posee tres grabaciones, repeticiones o patrones de covariables asociados, con 44 variables cada uno. De entre las variables consideradas, algunas pudieron conformar grupos debido a que se encontraban asociadas a una misma naturaleza de la característica sonora que medían, e.g. tono,

amplitud, ruido, frecuencia, etcétera. Por tal motivo los grupos poseían internamente altas correlaciones. Los grupos formados son Jitter, Shimmer, HNR, MFCC y Delta; entre los cuales se encuentran 40 de las 44 variables acústicas de la muestra.

5.4.1. Componentes principales comunes

Dado que la dimensión a menudo representa un problema en términos de la interpretación o visualización de los resultados, es necesario recurrir a herramientas que permitan reducirla. Una primera opción pudiera ser el ACP, sin embargo, más que seleccionar las componentes que mejor conserven la variabilidad de todas las variables de la muestra, el fin es seleccionar las componentes principales de cada grupo, tomando en cuenta que cada individuo tiene 3 repeticiones, por esto resultó conveniente utilizar el ACPC para recuperar los componentes que fuesen comunes a las tres repeticiones.

En el grupo Jitter se seleccionó un componente principal que conserva el 99.99 % de la variabilidad de los datos del grupo; en el grupo Shimmer se seleccionó también una componente, y conserva el 99.9 % de la variabilidad; en el grupo HNR se seleccionó un componente que aporta el 98.4 % de la variabilidad del grupo. En el grupo MFCC se seleccionaron dos componentes que permiten observar más del 80 % de la variabilidad original de los datos; mientras que en el grupo Delta, al seleccionar dos componentes principales sólo fue posible conservar el 78 % de la variabilidad. Lo anterior se debe a que las correlaciones en los grupos MFCC y Delta son más bajas que en los grupos Jitter, Shimmer y HNR, siendo particularmente menores en el grupo Delta, lo que se tradujo en una menor variabilidad aportada por los componentes seleccionados. La justificación para no seleccionar más componentes en los grupos MFCC y Delta es el principio de parsimonia, en aras de la dimensión de la submuestra final.

Cada uno de los componentes obtenidos a través del ACPC se considera como un índice, junto con las variables restantes, correspondientes a las medidas acústicas no lineales, que no pertenecían a alguno de los grupos. Del total inicial de las 40 variables que conformaban los grupos, se seleccionaron solamente 7 índices a través del ACPC.

De esta manera se conformaron los biomarcadores que permiten discriminar entre una persona con EP y otra sin la enfermedad. Los biomarcadores contienen los índices: Sexo, CPC-Jitter, CPC-Shimmer, CPC-HNR, RPDE, DFA, PPE, GNE, CPC1-MFCC, CPC2-MFCC, CPC1-Delta y CPC2-Delta. De este modo la muestra se redujo de un total de 45 variables a 12 índices asociados al “Estado” del individuo, donde el valor “1” representa a una persona con EP y “0” representa a una persona sin ella.

5.4.2. Regresión logística

El modelo lineal generalizado asociado a una variable dicotómica es el modelo de regresión logística

$$\pi(\mathbf{X}) = \frac{\exp(\beta^T \mathbf{X})}{1 + \exp(\beta^T \mathbf{X})},$$

donde $\pi(\mathbf{X})$ representa la probabilidad de tener la EP dado el conjunto de covariables $\mathbf{X}$ en el modelo lineal

$$\begin{aligned} \beta^T \mathbf{X} \;=\; & 67.9352 - 5.3129 \cdot \textbf{Sexo} + 21.5250 \cdot \textbf{CPC-Jitter} \\ & - 1.5649 \cdot \textbf{CPC-Shimmer} + 0.0868 \cdot \textbf{CPC-HNR} \\ & - 79.7407 \cdot \textbf{RPDE} - 24.8897 \cdot \textbf{DFA} + 0.4965 \cdot \textbf{PPE} \\ & + 8.6435 \cdot \textbf{GNE} + 6.6483 \cdot \textbf{CPC1-MFCC} \\ & - 14.7472 \cdot \textbf{CPC2-MFCC} + 0.2913 \cdot \textbf{CPC1-Delta} \\ & - 3.8843 \cdot \textbf{CPC2-Delta}, \end{aligned}$$

obtenido por el método de estimación de máxima verosimilitud sobre los datos de la muestra.

A pesar de haber examinado el ajuste del modelo, ninguno de los índices de los biomarcadores fue eliminado en el modelo de regresión logística dado que el propósito de este trabajo es encontrar un modelo parsimonioso utilizando el ACPC, por esto tampoco se utilizaron métodos de selección o eliminación de variables sobre el modelo de regresión.

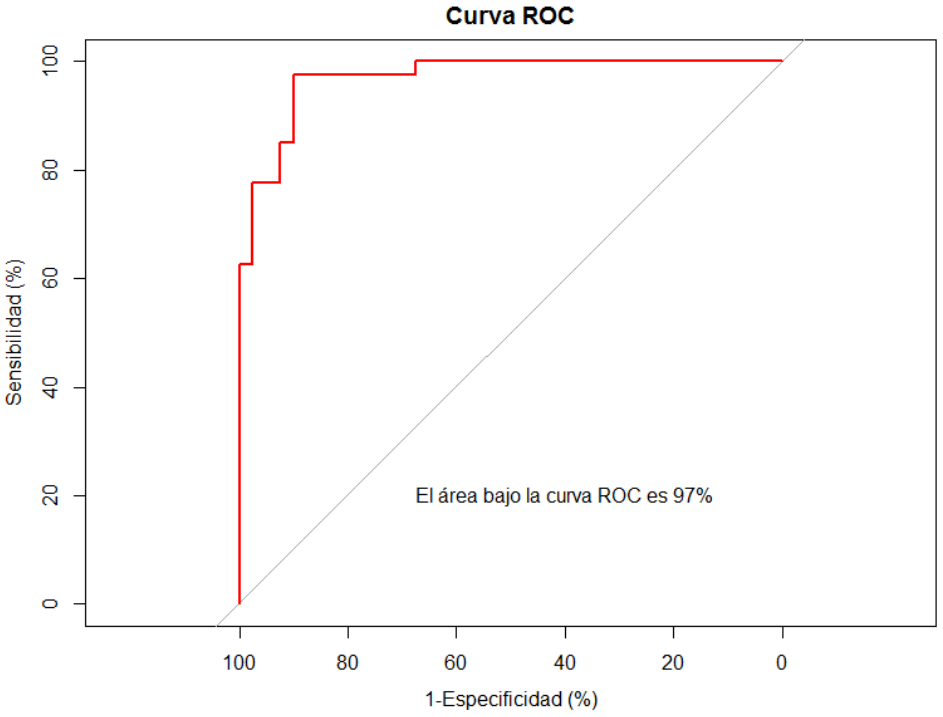
Algunas de las pruebas de bondad de ajuste del modelo anteriormente señalado, como la devianza y el AIC, arrojan poca información para conocer el grado de adecuamiento que tiene el modelo sobre los conjuntos de covariables de la muestra para determinar la respuesta observada puesto que sólo existe un modelo ajustado.

El análisis de la prueba ji-cuadrada de Pearson fue excluido de la sección correspondiente debido a que la prueba tiende a presentar errores cuando el tamaño de la muestra es grande y/o cuando los datos provienen de variables aleatorias continuas (como es el caso de la mayoría de las variables de la muestra).

No obstante, el modelo es bueno, puesto que en el análisis de los residuos, en la prueba de la devianza residual con hipótesis nula "El modelo está correctamente ajustado", la hipótesis nula no se rechaza. De modo que no existe evidencia estadística suficiente para rechazar que el modelo presenta un buen ajuste a los datos.

En la prueba de Hosmer-Lemeshow con hipótesis nula igual a la de la prueba de la devianza residual, la hipótesis nula tampoco se rechaza. De acuerdo con esta prueba, tampoco existe evidencia estadística suficiente para rechazar que el modelo presenta un buen ajuste a los datos. De hecho, esto es visible en los siguientes resultados.

El modelo resultante permite estimar el estado de las personas respecto a la EP, de manera tal que se obtienen 36 verdaderos positivos, 36 verdaderos negativos, 4 falsos positivos y 4 falsos negativos; una sensibilidad y especificidad de 90 %, un riesgo relativo igual a 9, una razón de momios igual a 81 y un área bajo la curva ROC de 97 %, determinando al método de diagnóstico como un método con discriminación sobresaliente.



El área bajo la curva ROC determina el método de diagnóstico como sobresaliente.

5.5. Validación cruzada

La *validación cruzada* es un método que sirve para evaluar el rendimiento de un modelo ante predicciones de datos que no han sido contemplados en el ajuste del modelo.

Un posible planteamiento de dicho método para la base de datos de este trabajo es seleccionar aleatoriamente 60 individuos de la muestra, 30 del grupo de personas con EP y 30 del grupo sin EP, y remover los 20 conjuntos de covariables restantes. Realizar el ACPC con los 60 sujetos seleccionados y sus correspondientes covariables, añadir las mediciones no lineales y ajustarles un modelo de regresión logística. Con los coeficientes ajustados obtenidos, conocer la probabilidad de tener la EP, utilizando los datos de

los 20 conjuntos de covariables que fueron apartados inicialmente, con la finalidad de comparar los resultados obtenidos con los valores observados en la variable "Estado". Por último se obtienen también la sensibilidad y especificidad, el riesgo relativo, la razón de momios y el área bajo la curva ROC, predichos por el modelo de regresión logística. Este procedimiento se realiza cien veces de manera automática y sólo se reportan las medias de cada una de las características mencionadas.

5.5.1. Componentes principales comunes

Para cada una de las iteraciones fue seleccionada la primer componente del grupo Jitter en la primera, segunda y tercera repetición. El primer CPC se denominó como índice CPC-Jitter. En el grupo Shimmer fue seleccionada la primer componente, el índice obtenido fue nombrado CPC-Shimmer. Análogamente fue seleccionada la primer componente del grupo HNR, CPC-HNR. En el grupo MFCC fueron seleccionadas las dos primeras componentes, los índices obtenidos fueron llamados CPC1-MFCC y CPC2-MFCC, respectivamente. En el grupo Delta se escogieron los primeros dos componentes, CPC1-Delta y CPC2-Delta, en ese mismo orden.

Incorporando las mediciones no lineales de las variables aisladas, la muestra de biomarcadores contiene los siguientes índices: Sexo, CPC-Jitter, CPC-Shimmer, CPC-HNR, RPDE, DFA, PPE, GNE, CPC1-MFCC, CPC2-MFCC, CPC1-Delta y CPC2-Delta.

5.5.2. Regresión logística

Con el modelo de regresión logística ajustado a los 60 conjuntos de covariables, intentamos predecir el estado de los 20 individuos restantes, por ello, tomando los coeficientes ajustados, se utilizan los valores de los 20 en cada índice.

La probabilidad $\pi(\mathbf{X})$ se obtiene como en (4.1) y con ella se obtienen las probabilidades de tener un diagnóstico positivo (1) o negativo (0) para la enfermedad de Parkinson. Dado que estas probabilidades no necesariamente son cero o uno, se asigna uno si la probabilidad estimada de presentar la respuesta fue mayor a 0.5, y cero en caso contrario. Una vez obtenida la probabilidad bajo el criterio anterior, se comparan los valores observados y los ajustados para obtener los verdaderos positivos, verdaderos negativos, falsos positivos, falsos negativos, la sensibilidad y especificidad, el riesgo relativo, la razón de momios y el área bajo la curva ROC para determinar la calidad del método de discriminación.

Dado que este procedimiento se realiza cien veces, con muestras de 60 individuos seleccionadas aleatoriamente cada vez, las cantidades presentadas como parte de la validación son únicamente los promedios de todas las iteraciones.

Las medias en la estimación del estado de las personas respecto a la EP, es la siguiente: 8 verdaderos positivos, 8 verdaderos negativos, 2 falsos positivos y 2 falsos negativos; una sensibilidad de 80 % y una especificidad de 78 %, un riesgo relativo igual a 3.68, una razón de momios igual a 14.44 y un área bajo la curva ROC de 86.06 %, determinando al método de diagnóstico como un método con discriminación excelente.

Capítulo 6

Conclusiones

Aplicación

Tras el análisis de la base de datos se conformaron cinco grupos en los cuales se agrupaban 40 de las 46 variables de la muestra, conforme a la naturaleza de las características acústicas que fueron medidas; el resto (a excepción de la variable "Sexo", y la variable "Estado", la cual fue considerada como la variable respuesta) fueron consideradas como mediciones aisladas. Con el análisis de los componentes principales comunes para cada uno de los cinco grupos formados, se seleccionaron siete componentes principales en total, los mismos que ofrecían la mejor reconstrucción de los datos originales.

Dichos componentes principales fueron considerados como índices junto con las mediciones aisladas no lineales, de modo que a partir de ellos se construyó un modelo de regresión logística. Los coeficientes del modelo de regresión en combinación con cada uno de los conjuntos de covariables de los individuos dentro del estudio, permitió obtener las probabilidades únicas para cada individuo de tener la EP. Para las probabilidades se asignó uno si la probabilidad estimada de presentar la respuesta fue mayor a 0.5, y cero en caso contrario, para hacerlas comparables con los datos categóricos de la variable "Estado".

Después de tal comparación, es observable que el modelo denotó con un diagnóstico "Positivo" a 36 de las 40 personas que tienen EP, i.e. la prueba tiene una sensibilidad del 90 %. Por el contrario, el modelo denotó con un diagnóstico "Negativo" a 36 de las 40 personas del grupo control, i.e. la prueba tiene una especificidad del 90 %. Por tanto, es posible considerar el modelo como un método de diagnóstico bueno, tomando en cuenta también sus limitaciones. Entre ellas, que el modelo ha servido en particular para la población en estudio y que, aunque sin duda sirve para el análisis de nuevos registros

vocales, ha de tenerse en consideración, que a medida que se añadan más individuos a la muestra, el modelo debería replantearse y construirse nuevamente para maximizar la probabilidad de obtener un diagnóstico correcto en los nuevos conjuntos de covariables agregados a la muestra.

Otros métodos

Algunos otros métodos útiles para clasificar datos son, principalmente, el método de los k-vecinos más cercanos (*k-nearest neighbors*), la máquina de vectores de soporte (*support vector machine*), regresión binaria no lineal (*nonlinear binary regression*), análisis de discriminante (*discriminant function analysis*) y redes neuronales (*neural networks*).

Medidas repetidas

En este trabajo se ha abordado el análisis de un conjunto de datos con medidas repetidas en las covariables y se ha determinado promediar los valores de tales repeticiones en cada uno de los índices de los biomarcadores de cada individuo. La razón de proceder de tal manera es que, si bien existen modelos para medidas repetidas, e.g. modelos mixtos (Berridge y Crouchley, 2011) o modelos multinivel (Goldstein, 2010), sería necesario que la variable que se toma como variable respuesta fuese también una medida repetida y, sin embargo, en el problema planteado la variable respuesta no es una medida repetida, sino una cuyo valor permanece invariante durante las repeticiones debido a que se trata del estado del individuo con respecto a la EP; de manera que en este trabajo, sólo las covariables han sido medidas con repeticiones.

En la representación de los modelos mixtos, las variables regresoras son asociadas a una variable respuesta que se mide a través de una variable binaria, sin embargo, la variable respuesta en sí es una variable con múltiples niveles [2]. En el caso de los modelos multinivel, la variable respuesta puede tomar una distribución binomial o Poisson, el primer caso cuando la respuesta es una proporción, y el segundo cuando es un conteo; de cualquier modo, la respuesta es contada como una medida repetida [5]. Por ello, la solución propuesta a este problema fue el promedio de las medidas observadas en las repeticiones.

Apéndice A

Anexos

Ortogonalidad

$\mathbf{A} = [a_1, a_2, \ldots, a_p]$ es una *matriz ortogonal* si $\mathbf{A}$ es invertible y

1. $\mathbf{A}^{-1} = \mathbf{A}^T$, o
2. $a_i \cdot a_j = 0$ si $i \neq j$, y $a_i \cdot a_j = 1$ si $i = j$,

con cada vector columna $a_i \in \mathbb{R}^q$, $i, j = 1, \ldots, p$, para alguna $q \in \mathbb{N}$.

Una consecuencia inmediata del primer punto es que el producto $\mathbf{A}\mathbf{A}^T$ da como resultado la *matriz identidad* $\mathbb{I}_q$.

Distribuciones discretas de la familia exponencial natural

CUADRO A.1: Distribuciones discretas de la familia exponencial natural.

Distribución	$a_0(\theta)$	$b_0(y)$	$Q(\theta)$
$Bin(n, \theta)$	$\exp\left[-n \log(1 + e^{Q(\theta)})\right]$	$\binom{n}{y}$	$\log\left(\frac{\theta}{1-\theta}\right)$
$Poisson(\theta)$	$e^{-\theta}$	$\frac{1}{y!}$	$\log(\theta)$
$BinNeg(k, \theta)$	$k \ \log\left(\frac{1 - e^{Q(\theta)}}{e^{Q(\theta)}}\right)$	$\binom{y-1}{k-1}$	$\log(1-\theta)$

Funciones liga

CUADRO A.2: Funciones liga más comunes para modelos de respuesta binaria [11].

Liga	Nombre común	Función liga
Función logit	*logit*	$\log\{\pi/(1-\pi)\}$
Función normal inversa	*probit*	$\Phi^{-1}(\pi)$
Función log log	*loglog*	$-\log\{-\log(\pi)\}$
Función log log complementaria	*cloglog*	$\log\{-\log(1-\pi)\}$

Método delta

El método delta es un método para obtener la varianza asintótica de una estadística de prueba.

Sea $f(\theta)$ una función de una estadística θ, entonces, el método delta asegura que el error estándar de $f(\theta)$ es

$$se(f(\theta)) = \left|\frac{d\ f(\theta)}{d\theta}\right| \cdot se(\theta),$$

donde *se* significa error estándar.

En el caso de una tabla de contingencia como la que se presenta a continuación, al tratar de obtener la varianza de una estadística θ no es posible utilizar el caso univariado del método delta, de modo que utilizamos su versión multivariada.

Considere la siguiente tabla de contigencia,

	Columna		
Renglón	1	2	Total
1	n_{11}	n_{12}	$n_{1\bullet}$
2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet} = n$

Suponga que la estadística θ es la estimación de RP, i.e.

$$\theta = \widehat{RP} = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

entonces la versión multivariada del método delta es

$$\mathbb{V}(\theta) \approx \nabla f(n_{11}, n_{12}, n_{21}, n_{22}) \cdot Cov(n_{11}, n_{12}, n_{21}, n_{22}) \cdot \nabla f(n_{11}, n_{12}, n_{21}, n_{22})^T,$$

donde $\nabla f(n_{11}, n_{12}, n_{21}, n_{22}) = \left(\frac{\partial f}{\partial n_{11}}, \frac{\partial f}{\partial n_{12}}, \frac{\partial f}{\partial n_{21}}, \frac{\partial f}{\partial n_{22}}\right)$.

Como deseamos estimar la varianza del logaritmo de $\widehat{RP}$, i.e.

$$\mathbb{V}\left(\log\left(\widehat{RP}\right)\right) = \mathbb{V}\left(\log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right)\right),$$

de acuerdo al método delta necesitamos definir una función f de la estadística θ, así que definimos

$$f(n_{11}, n_{12}, n_{21}, n_{22}) = \log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right) = (\log n_{11} - \log n_{12} - \log n_{21} + \log n_{22}),$$

por lo tanto

$$\nabla f(n_{11}, n_{12}, n_{21}, n_{22}) = \left(\frac{1}{n_{11}}, \frac{-1}{n_{12}}, \frac{-1}{n_{21}}, \frac{1}{n_{22}}\right).$$

La matriz de covarianzas de una distribución multinomial con 4 categorías es

$$\Sigma = \begin{bmatrix} n_{11}(1-n_{11}) & -n_{12}n_{11} & -n_{21}n_{11} & -n_{22}n_{11} \\ -n_{11}n_{12} & n_{12}(1-n_{12}) & -n_{21}n_{12} & -n_{22}n_{12} \\ -n_{11}n_{21} & -n_{12}n_{21} & n_{21}(1-n_{21}) & -n_{22}n_{21} \\ -n_{11}n_{22} & -n_{12}n_{22} & -n_{21}n_{22} & n_{22}(1-n_{22}) \end{bmatrix}$$

Entonces $\mathbb{V}\left(\log\left(\widehat{RP}\right)\right)$

$$\approx \begin{bmatrix} \frac{1}{n_{11}} \\ \frac{-1}{n_{12}} \\ \frac{-1}{n_{21}} \\ \frac{1}{n_{22}} \end{bmatrix}^T \begin{bmatrix} n_{11}(1-n_{11}) & -n_{12}n_{11} & -n_{21}n_{11} & -n_{22}n_{11} \\ -n_{11}n_{12} & n_{12}(1-n_{12}) & -n_{21}n_{12} & -n_{22}n_{12} \\ -n_{11}n_{21} & -n_{12}n_{21} & n_{21}(1-n_{21}) & -n_{22}n_{21} \\ -n_{11}n_{22} & -n_{12}n_{22} & -n_{21}n_{22} & n_{22}(1-n_{22}) \end{bmatrix} \begin{bmatrix} \frac{1}{n_{11}} \\ \frac{-1}{n_{12}} \\ \frac{-1}{n_{21}} \\ \frac{1}{n_{22}} \end{bmatrix}$$

$$= \begin{bmatrix} (1-n_{11}) + n_{11} + n_{11} - n_{11} \\ -n_{12} - (1-n_{12}) + n_{12} - n_{12} \\ -n_{21} + n_{21} - (1-n_{21}) - n_{21} \\ -n_{22} + n_{22} + n_{22} + (1-n_{22}) \end{bmatrix}^T \begin{bmatrix} \frac{1}{n_{11}} \\ \frac{-1}{n_{12}} \\ \frac{-1}{n_{21}} \\ \frac{1}{n_{22}} \end{bmatrix} = \left[\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right].$$

Por lo tanto, de acuerdo al método delta

$$\mathbb{V}\left(\log\left(\widehat{RP}\right)\right) \approx \left[\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right].$$

Apéndice B

Código de R

```
# Se cargan los datos al programa:
Parkinson<-read.csv("E:/MedidasAcusticas.csv", header=T, dec=",",
                    + sep=";")
attach(Parkinson)
names(Parkinson)

# Creamos los 5 grupos, a partir de los cuales se obtendrán los CPC:
Jitter=cbind(Jitter_rel,Jitter_abs,Jitter_RAP,Jitter_PPQ)
Shimmer=cbind(Shim_loc,Shim_dB,Shim_APQ3,Shim_APQ5,Shi_APQ11)
HNR=cbind(HNR05,HNR15,HNR25,HNR35,HNR38)
MFCC=cbind(MFCC0,MFCC1,MFCC2,MFCC3,MFCC4,MFCC5,MFCC6,MFCC7,MFCC8,
           + MFCC9,MFCC10,MFCC11,MFCC12)
Delta=cbind(Delta0,Delta1,Delta2,Delta3,Delta4,Delta5,Delta6,Delta7,
           + Delta8,Delta9,Delta10,Delta11,Delta12)

# Obtenemos sus matrices de covarianzas y de correlaciones:
covJitter=cov(Jitter)
covShimmer=cov(Shimmer)
covHNR=cov(HNR)
covMFCC=cov(MFCC)
covDelta=cov(Delta)

corJitter=cor(Jitter)
corShimmer=cor(Shimmer)
corHNR=cor(HNR)
corMFCC=cor(MFCC)
```

```
corDelta=cor(Delta)

library(corrplot)
corrplot(corJitter, method = "color")
par(mfrow=c(1,2))
corrplot.mixed(corShimmer, lower = "number", upper = "color")
corrplot.mixed(corHNR, lower = "number", upper = "color")
par(mfrow=c(1,2))
corrplot.mixed(corMFCC, lower = "number", upper = "color")
corrplot.mixed(corDelta, lower = "number", upper = "color")

#***********************
# Repeticiones número 1
#***********************
# Se cargan los datos al programa:
Parkinson1<-read.csv("E:/Uno.csv", header=T, dec=",", sep=";")
attach(Parkinson1)
names(Parkinson1)

# Creamos los 5 grupos, a partir de los cuales se obtendrán los CPC:
X1Jitter=cbind(X1Jitter_rel,X1Jitter_abs,X1Jitter_RAP,X1Jitter_PPQ)
X1Shimmer=cbind(X1Shim_loc,X1Shim_dB,X1Shim_APQ3,X1Shim_APQ5,X1Shi_APQ11)
X1HNR=cbind(X1HNR05,X1HNR15,X1HNR25,X1HNR35,X1HNR38)
X1MFCC=cbind(X1MFCC0,X1MFCC1,X1MFCC2,X1MFCC3,X1MFCC4,X1MFCC5,X1MFCC6,
             + X1MFCC7,X1MFCC8,X1MFCC9,X1MFCC10,X1MFCC11,X1MFCC12)
X1Delta=cbind(X1Delta0,X1Delta1,X1Delta2,X1Delta3,X1Delta4,X1Delta5,
              + X1Delta6,X1Delta7,X1Delta8,X1Delta9,X1Delta10,X1Delta11,
              + X1Delta12)

# Obtenemos sus matrices de covarianzas y de correlaciones:
covX1Jitter=cov(X1Jitter)
covX1Shimmer=cov(X1Shimmer)
covX1HNR=cov(X1HNR)
covX1MFCC=cov(X1MFCC)
covX1Delta=cov(X1Delta)


#***********************
# Repeticiones número 2
```

```
#***********************
# Se cargan los datos al programa:
Parkinson2<-read.csv("E:/Dos.csv", header=T, dec=",", sep=";")
attach(Parkinson2)
names(Parkinson2)
# Creamos los 5 grupos, a partir de los cuales se obtendrán los CPC:
X2Jitter=cbind(X2Jitter_rel,X2Jitter_abs,X2Jitter_RAP,X2Jitter_PPQ)
X2Shimmer=cbind(X2Shim_loc,X2Shim_dB,X2Shim_APQ3,X2Shim_APQ5,X2Shi_APQ11)
X2HNR=cbind(X2HNR05,X2HNR15,X2HNR25,X2HNR35,X2HNR38)
X2MFCC=cbind(X2MFCC0,X2MFCC1,X2MFCC2,X2MFCC3,X2MFCC4,X2MFCC5,X2MFCC6,
             + X2MFCC7,X2MFCC8,X2MFCC9,X2MFCC10,X2MFCC11,X2MFCC12)
X2Delta=cbind(X2Delta0,X2Delta1,X2Delta2,X2Delta3,X2Delta4,X2Delta5,
             + X2Delta6,X2Delta7,X2Delta8,X2Delta9,X2Delta10,X2Delta11,
             + X2Delta12)

# Obtenemos sus matrices de covarianzas y de correlaciones:
covX2Jitter=cov(X2Jitter)
covX2Shimmer=cov(X2Shimmer)
covX2HNR=cov(X2HNR)
covX2MFCC=cov(X2MFCC)
covX2Delta=cov(X2Delta)

#***********************
# Repeticiones número 3
#***********************
# Se cargan los datos al programa:
Parkinson3<-read.csv("E:/Tres.csv", header=T, dec=",", sep=";")
attach(Parkinson3)
names(Parkinson3)

# Creamos los 5 grupos, a partir de los cuales se obtendrán los CPC:
X3Jitter=cbind(X3Jitter_rel,X3Jitter_abs,X3Jitter_RAP,X3Jitter_PPQ)
X3Shimmer=cbind(X3Shim_loc,X3Shim_dB,X3Shim_APQ3,X3Shim_APQ5,X3Shi_APQ11)
X3HNR=cbind(X3HNR05,X3HNR15,X3HNR25,X3HNR35,X3HNR38)
X3MFCC=cbind(X3MFCC0,X3MFCC1,X3MFCC2,X3MFCC3,X3MFCC4,X3MFCC5,X3MFCC6,
             + X3MFCC7,X3MFCC8,X3MFCC9,X3MFCC10,X3MFCC11,X3MFCC12)
X3Delta=cbind(X3Delta0,X3Delta1,X3Delta2,X3Delta3,X3Delta4,X3Delta5,
             + X3Delta6,X3Delta7,X3Delta8,X3Delta9,X3Delta10,X3Delta11,
             + X3Delta12)
```

```
# Obtenemos sus matrices de covarianzas y de correlaciones:
covX3Jitter=cov(X3Jitter)
covX3Shimmer=cov(X3Shimmer)
covX3HNR=cov(X3HNR)
covX3MFCC=cov(X3MFCC)
covX3Delta=cov(X3Delta)

# ******************************************
# Componentes principales para el grupo Jitter:
# ******************************************
AJ<-cbind(covX1Jitter,covX2Jitter)
BJ<-cbind(AJ,covX3Jitter)
dim(BJ)<-c(4,4,3)

library("cpca", lib.loc="C:/Program Files/R/R-3.2.4revised/library")
modJ<-cpc(BJ)
modJ
cumsum(modJ$D[,1])/sum(modJ$D[,1])
cumsum(modJ$D[,2])/sum(modJ$D[,2])
cumsum(modJ$D[,3])/sum(modJ$D[,3])

#Scores solo con la primer componente:
cpcJ1 = X1Jitter%*%modJ$CPC[,1]
cpcJ2 = X2Jitter%*%modJ$CPC[,1]
cpcJ3 = X3Jitter%*%modJ$CPC[,1]
cpcJ = rowMeans(cbind(cpcJ1,cpcJ2,cpcJ3))
# ******************************************

# ******************************************
# Componentes principales para el grupo Shimmer:
# ******************************************
AS<-cbind(covX1Shimmer,covX2Shimmer)
BS<-cbind(AS,covX3Shimmer)
dim(BS)<-c(5,5,3)

library("cpca", lib.loc="C:/Program Files/R/R-3.2.4revised/library")
modS<-cpc(BS)
modS
```

```
cumsum(modS$D[,1])/sum(modS$D[,1])
cumsum(modS$D[,2])/sum(modS$D[,2])
cumsum(modS$D[,3])/sum(modS$D[,3])

#Scores solo con la primer componente:
cpcS1 = X1Shimmer%*%modS$CPC[,1]
cpcS2 = X2Shimmer%*%modS$CPC[,1]
cpcS3 = X3Shimmer%*%modS$CPC[,1]
cpcS = rowMeans(cbind(cpcS1,cpcS2,cpcS3))
# ****************************************

# ****************************************
# Componentes principales para el grupo HNR:
# ****************************************
AH<-cbind(covX1HNR,covX2HNR)
BH<-cbind(AH,covX3HNR)
dim(BH)<-c(5,5,3)

library("cpca", lib.loc="C:/Program Files/R/R-3.2.4revised/library")
modH<-cpc(BH)
modH
cumsum(modH$D[,1])/sum(modH$D[,1])
cumsum(modH$D[,2])/sum(modH$D[,2])
cumsum(modH$D[,3])/sum(modH$D[,3])

#Scores solo con la primer componente:
cpcH1 = X1HNR%*%modH$CPC[,1]
cpcH2 = X2HNR%*%modH$CPC[,1]
cpcH3 = X3HNR%*%modH$CPC[,1]
cpcH = rowMeans(cbind(cpcH1,cpcH2,cpcH3))
# ****************************************

# ****************************************
# Componentes principales para el grupo MFCC:
# ****************************************
AM<-cbind(covX1MFCC,covX2MFCC)
BM<-cbind(AM,covX3MFCC)
dim(BM)<-c(13,13,3)
```

```
library("cpca", lib.loc="C:/Program Files/R/R-3.2.4revised/library")
modM<-cpc(BM)
modM
cumsum(modM$D[,1])/sum(modM$D[,1])
cumsum(modM$D[,2])/sum(modM$D[,2])
cumsum(modM$D[,3])/sum(modM$D[,3])

#Scores con las primeras dos componentes:
cpcM11 = X1MFCC%*%modM$CPC[,1]
cpcM12 = X2MFCC%*%modM$CPC[,1]
cpcM13 = X3MFCC%*%modM$CPC[,1]
cpcM1 = rowMeans(cbind(cpcM11,cpcM12,cpcM13))

cpcM21 = X1MFCC%*%modM$CPC[,2]
cpcM22 = X2MFCC%*%modM$CPC[,2]
cpcM23 = X3MFCC%*%modM$CPC[,2]
cpcM2 = rowMeans(cbind(cpcM21,cpcM22,cpcM23))

#Plot#
install.packages("ggplot2")
library(ggplot2)
install.packages("gridExtra")
library(gridExtra)
install.packages("grid")
library(grid)
install.packages("Rmisc")
library(Rmisc)
#Combinando plots
q1=quickplot(cpcM11,cpcM21, main="Primera repetición",
            + xlab="CPC1-MFCC", ylab="CPC2-MFCC")
q2=quickplot(cpcM12,cpcM22, main="Segunda repetición",
            + xlab="CPC1-MFCC", ylab="CPC2-MFCC")
q3=quickplot(cpcM13,cpcM23, main="Tercera repetición",
            + xlab="CPC1-MFCC", ylab="CPC2-MFCC")
q4=quickplot(cpcM1,cpcM2, main="Promedio",
            + xlab="CPC1-MFCC", ylab="CPC2-MFCC")
multiplot(q1, q3, q2, q4, cols=2)
# ************************************************
```

```
# ******************************************
# Componentes principales para el grupo Delta:
# ******************************************
AD<-cbind(covX1Delta,covX2Delta)
BD<-cbind(AD,covX3Delta)
dim(BD)<-c(13,13,3)

library("cpca", lib.loc="C:/Program Files/R/R-3.2.4/library")
modD<-cpc(BD)
modD
cumsum(modD$D[,1])/sum(modD$D[,1])
cumsum(modD$D[,2])/sum(modD$D[,2])
cumsum(modD$D[,3])/sum(modD$D[,3])

#Scores con las primeras dos componentes:
cpcD11 = X1Delta%*%modD$CPC[,1]
cpcD12 = X2Delta%*%modD$CPC[,1]
cpcD13 = X3Delta%*%modD$CPC[,1]
cpcD1 = rowMeans(cbind(cpcD11,cpcD12,cpcD13))

cpcD21 = X1Delta%*%modD$CPC[,2]
cpcD22 = X2Delta%*%modD$CPC[,2]
cpcD23 = X3Delta%*%modD$CPC[,2]
cpcD2 = rowMeans(cbind(cpcD21,cpcD22,cpcD23))
#Plots
q1=quickplot(cpcD11,cpcD21, main="Primera repetición",
xlab="CPC1-Delta", ylab="CPC2-Delta")
q2=quickplot(cpcD12,cpcD22, main="Segunda repetición",
xlab="CPC1-Delta", ylab="CPC2-Delta")
q3=quickplot(cpcD13,cpcD23, main="Tercera repetición",
xlab="CPC1-Delta", ylab="CPC2-Delta")
q4=quickplot(cpcD1,cpcD2, main="Promedio",
xlab="CPC1-Delta", ylab="CPC2-Delta")
multiplot(q1, q3, q2, q4, cols=2)
# ******************************************

#*********************
# Regresión logística:
#*********************
```

```
# Se cargan los datos al programa:
ParkinP<-read.csv("E:/ParkinP.csv", header=T)
attach(ParkinP)
names(ParkinP)
ParkinP=ParkinP[,-c(3:5,10:21)]
ParkinP=cbind(ParkinP,cpcJ,cpcS,cpcH,cpcM1,cpcM2,cpcD1,cpcD2)
logit<-glm(XPEstado ~ XPSexo+cpcJ+cpcS+cpcH+XPRPDE+XPDFA+XPPPE
                    +XPGNE+cpcM1+cpcM2+cpcD1+cpcD2,
                    +family = "binomial")
logit
summary(logit)

# Intervalos de confianza:
confint(logit)

#******************
#Ajuste del modelo:
#******************
#cociente de verosimilitudes:
logit$null.deviance-logit$deviance

#Prueba de Wald:
#instalamos el paquete "aod".
install.packages("aod")
library(aod)
# Wald test para cada coeficiente de la regresión.
for(i in 1:13)
{
  j=wald.test(b = coef(logit), Sigma= vcov(logit), Terms = i)
  print(j)
}

#*****************
#Bondad de ajuste:
#*****************
#Devianza:
logit$deviance
logLik(logit)
```

```
#*************************
# Análisis de los residuos:
#*************************
#Devianza residual total:
sum((residuals(logit))^2)
#Comparación con devianza del modelo:
logit$deviance
#X2 con 67 g.l. y a=0.05:
qchisq(.95, df=67)

#Prueba de Hosmer-Lemeshow:
install.packages("ResourceSelection")
library(ResourceSelection)
hoslem.test(Estado, fitted(logit))
#X2 con 2 g.l. y a=0.05:
qchisq(0.95,df=2)
#[1] 5.991465
```

Bibliografía

[1] Agresti, Alan (2002). *Categorical Data Analysis*. John Wiley & Sons, USA.

[2] Berridge, Damon M. y Crouchley, Robert (2011). *Multivariate Generalized Linear Mixed Models Using R*. CRC Press, USA.

[3] Dorfsman, L. K. (2006). Componentes principales comunes: un método para la comparación de matrices de varianza y covarianza. Universidad Autónoma de México.

[4] Flury, Bernhard (1988). *Common Principal Components Analysis and Related Multivariate Models*. John Wiley & Sons, USA.

[5] Goldstein, Harvey (2010). *Multilevel Statistical Models*. John Wiley & Sons, UK, fourth edition.

[6] Hanley, J. A.; y McNeil, B. J. (1982). The meaning and use of the area under a receriver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

[7] Harel, B.; Cannizzaro, M.; y Snyder, P. J. (2004). Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study. *Brain and Cognition*, 56:24–29.

[8] Hosmer, David W. y Lemeshow, Stanley (2000). *Applied Logistic Regression*. John Wiley & Sons, USA, second edition.

[9] Little, M. A.; McSharry, P. E.; Hunter, E. J.; Spielman, J. y Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022.

[10] McCullagh, Peter (2002). What is a statistical model? *The Annals of Statistics*, 30(5):1225–1310.

[11] McCullagh, Peter y Ashworth Nelder, John (1989). *Generalized Linear Models*. Chapman and Hall, USA, second edition.

[12] Naranjo, Lizbeth; Pérez, Carlos J.; Campos-Roca, Yolanda y Martín, Jacinto (2016). Addressing voice recording replications for Parkinson's disease detection. *Expert Systems With Applications*, 46:286–292.

[13] Rajput, M.; Rajput, A. y Rajput, A. H. (2007). *Epidemiology (chapter 2) in Handbook of Parkinson's disease*. Informa Healthcare, USA, fourth edition.

[14] Sapir, S.; Ramig, L.; Spielman, J.; y Fox, C. (2010). Formant centralization ratio (fcr): A proposal for a new acoustic measure of dysarthric speech. *Journal of Speech Language and Hearing Research*, 53:114–125.

[15] Tsanas, A.; Little, M. A.; McSharry, P. E. y Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893.

[16] Tsanas, A.; Little, M. A.; McSharry, P. E. y Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society Interface*, 8:842–855.

[17] Tsanas, Athanasios; Little, Max A.; McSharry, Patrick E.; Spielman, Jennifer y Ramig, Lorraine O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271.

[18] Van Den Eeden, S. K.; Tanner, C. M.; Bernstein, A. L.; Fross, R. D.; Leimpeter, A.; Bloch, D. A.; y Nelson, L. M. (2003). Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. *American Journal of Epidemiology*, 157:1015–1022.