Centro de Investigación en Matemáticas, A. C.

CIMAT

"Una aplicación de PLS no métrico para respuestas múltiples"

T E S I S

Para obtener el grado de:

Maestro en Ciencias
con Orientación en Probabilidad y Estadística

PRESENTA:

Act. Bryan Yaset Agüero Cruz

Director:

Dr. Graciela González Farías

Guanajuato, Gto, México.

Enero de 2014

Integrantes del Jurado

Presidente: Dr. Johan Van Horebek
Secretario: Dra. Belem Trejo Valdivia
Vocal y director de tesis: Dra. Graciela González Farías

Director:

Dra. Graciela González Farías

Sustentante:

Act. Bryan Yaset Agüero Cruz

Dedico este trabajo a mis padres, Miguel Ángel Agüero Sánchez y María Teresa Cruz Lizardi, quienes me han apoyado en todo momento y han formado a la persona que soy en día. A mis hermanos Mishel y Alan con quienes he compartido momentos inolvidables.

Agradecimientos

Me complace agradece a mi directora de tesis, a la Dra. Graciela María González Farías quien me brindó su conocimiento, su tiempo, su sinceridad y apoyo, por guiarme en este trabajo e impulsar en mi nuevos retos y en especial por su amistad y confianza durante la elaboración de la tesis y mi estancia en Monterrey.

A mis padres por apoyarme en en todo momento y brindarme sus consejos, por todo el cariño y amor que me brindaron durante toda mi vida; a mis profesores, el Dr. Rogelio Ramos Quiroga por todas sus enseñanzas durante mi estancia en CIMAT, por sus consejos y aportaciones a mi vida académica, profesional y personal, a mi tutor el Dr. Enrique Villa Diharce, al coordinador de la maestría el Dr. Joaquín Ortega Sánchez por sus consejos y sugerencias durante la maestría, al Dr. Antonio Costilla lector especial de la tesis y en especial por sus aportaciones a este trabajo, al Dr. Miguel Nakamura Savoy por compartirme su forma de pensar, sus ideales y sobre todo por cambiar mi perspectiva de las cosas; a mis amigos de la maestría quienes me apoyaron, me escucharon y creamos una gran familia durante la maestría.

A mis abuelitas Ma. Teresa Lizardi Robles y Rebeca Sánchez Chaparro y a todos y cada uno de mis tíos y familiares por el cariño y la comprensión que tienen hacia mi persona y a mi familia.

Para finalizar agradezco en especial al CONACYT, por la beca que me brindaron para realizar mis estudios de maestría, y por el apoyo brindando por el proyecto CONACYT: Ciencia Básica No. 105657 para poder realizar el viaje a Monterrey, sin los cuales no hubiera podido realizar este trabajo.

Resumen

La metodología estadística ha sido exitosamente usada en muchos tipos de problemas, el método de regresión más popular en la Quimiometría es el método de mínimos cuadrados parciales (PLS por sus siglas en inglés), este método fue propuesto por H. Wold en 1975 en su artículo titulado "*Soft Modeling by Latent Variables; the Non-linear Iterative Least Squares*". El método de PLS es usado para formar una relación entre las variables predictoras X y las variables predichas Y.

PLS se originó inicialmente como una técnica para datos en espacios continuos, a través de los años en casos donde la información es categórica u ordinal existen metodologías alternas. Entre estas metodologías no existe una que predomine sobre las otras, ya que se basan en diferentes enfoques de acuerdo al problema que se presente y esto dificulta su comparación. Dentro de las metodologías recientes se encontraron dos enfoques diferentes; el enfoque de "*path analysis*" [4] y el enfoque "*no métrico de PLS*" [14]. El primero posee un enfoque en términos de variables latentes y ecuaciones estructurales. El segundo posee un enfoque en un reescalamiento óptimo multidimensional, y es el que se aborda en esta tesis.

El estudio que motiva esta tesis forma parte de otro proyecto, dirigido a estudiar genéticamente a una población expuesta al humo del cigarrillo. Se estudiaron sustancias en líquidos corporales y se creó un cuestionario para medir las características de los fumadores activos y pasivos, que pudieran estar relacionadas con las concentraciones de las sustancias. El objetivo de la tesis es encontrar esa relación en caso de que existiera, con dicha relación se puede obtener un modelo de predicción fuera de muestra a nivel individual y con un algoritmo de aprendizaje.

En este trabajo se muestra la teoría básica de PLS en el capítulo 1, también se explican las ventajas de PLS y la forma de calcular PLS univariado y multivariado. Seguido de lo anterior en el capitulo 2, se presenta el enfoque de Mínimos Cuadrados Parciales no métricos, y los algoritmos para calcularlo. Adicional a esto, se presenta la parte fundamental del Bootstrap para crear intervalos de confianza bootstrap para los coeficientes de regresión, afín de poder presentar las capacidades predictivas del método. En el capitulo 3 se describe el ejemplo del tabaquismo y la base de datos relacionada, más aún se realiza un análisis descriptivo y estadístico de la muestra haciendo uso de las tablas de contingencia y del análisis de correspondencia. En el capítulo 4, se usa el enfoque de regresión PLS no métrico en el ejemplo del tabaquismo, es decir, cuando las variables predictoras son variables ordinales.

Índice general

Capítulo 1

Marco teórico de PLS

La metodología estadística ha sido exitosamente usada en muchos tipos de problemas, especialmente en la industria química. A finales del siglo XX la Quimiometría[1] emergió con el objetivo de analizar los datos observados de diferentes fenómenos, esta información observacional se enfocó en variables medibles con pocas observaciones, las cuales tendieron a servir para caracterizar los fenómenos de interés. El método de regresión más popular en la Quimiometría es mínimos cuadrados parciales (PLS por sus siglas en inglés), este método fue propuesto por H. Wold en 1975 en su artículo titulado "*Soft Modeling by Latent Variables; the Non-linear Iterative Least Squares*".

PLS es un método que se ha usado en diversas situaciones y tipos de estudios, a pesar de ello es un método que causa polémica, por ejemplo, una de las preguntas que naturalmente surgen con respecto a este tema es, ¿Por qué se espera que regresión PLS sea mejor que regresión lineal múltiple, regresión Ridge o cualquier otra técnica de regresión?. Otro factor importante es acerca del número de componentes PLS a considerar en el modelo, ya que probablemente se desea reducir la dimensionalidad de las variables predictoras sin perder el poder predictivo del modelo original. Una característica del método PLS es la forma en que se realiza el análisis de los datos asociados, es decir, de forma secuencial y tomando los objetos X y Y en parejas.

El método de PLS puede ser usado de forma univariada o multivariada, para formar la relación entre las variables predictoras X y las variables predichas Y. PLS construye nuevas variables predictoras, las cuales reciben diversos nombres, algunos de ellos son: factores, variables latentes o componentes, entre otros. Cada uno de estos componentes es una combinación lineal de las variables predictoras X. La idea en general es formar combinaciones de las X que predigan a las combinaciones de las Y.

[1] La Quimiometría es una disciplina de la química que se enfoca en la aplicación de métodos matemáticos y estadísticos para diseñar o seleccionar procedimientos de medida y experimentos óptimos para proporcionar la mayor información mediante el análisis de datos químicos.

1.1. Regresión PLS

El enfoque de regresión PLS (PLSR) predice una o varias variables dependientes a la vez, como una combinación lineal de un conjunto de variables predictoras, o como una combinación lineal de variables latentes $\mathbf{t} = (\mathbf{t}_{(1)1}, ..., \mathbf{t}_{(1)h}, ..., \mathbf{t}_{(1)H})$ [3]. Al mismo tiempo, también es una potente herramienta de visualización, ya que las variables latentes componen un subespacio de menor dimensión en comparación con el espacio formado por las variables predictoras originales, y éste subespacio es útil para explicar las variables respuestas. Con éste procedimiento se maximiza la información de las variables predictoras originales para predecir la mejor información de las variables predichas.

1.1.1. Algoritmo PLS

Sea X una matriz centrada de $N \times P_1$ y Y una matriz centrada de $N \times P_2$, y sin ningún otro supuesto adicional, el algoritmo es el siguiente

Paso I: Se selecciona un vector inicial cualquiera $t_{(2)i}$ (primer componente de Y)

Paso II: $w_1 = \frac{X^T t_{(2)i}}{t_{(2)i}^T t_{(1)i}}$

Paso III: $w_1 = \frac{w_1}{||w_1||}$

Paso IV: $t_{(1)i} = X w_1$

Paso V: $c_1 = \frac{Y^T t_{(1)i}}{t_{(1)i}^T t_{(1)i}}$

Paso VI: $c_1 = \frac{c_1}{||c_1||}$

Paso VII: $t_{(2)i} = Y^T c_1 / c_1^T c_1$

Paso VIII: Repetir los pasos del **II** al **VII** hasta la convergencia de $t_{(2)i}$ (e implícitamente la convergencia de c_1)

Paso IX: $p = X^T t_{(1)i} / (t_{(1)i}^T t_{(1)i})$

Paso X: $q = Y^T t_{(2)i} / (t_{(2)i}^T t_{(2)i})$

Paso XI: $b = t_{(2)i}^T t_{(1)i} / (t_{(1)i}^T t_{(1)i})$

Paso XII: $X = X - t_{(1)i} p^T$ y $Y = Y - b t_{(1)i} c^t$

donde c_1 y q son escalares y $t_{(2)i}$ es proporcional a Y cuando se tiene el caso univariado. Intrínsecamente el algoritmo calcula los vectores propios a través del método de potencia, el cual es un algoritmo numérico para aproximar los valores y vectores propios de una matriz. Una vez que se calcularon los primeros componentes $t_{(2)i}$ y $t_{(1)i}$, se procede a calcular los

residuos. Luego se efectúa el algoritmo anterior pero con los residuos recién calculados, esto es para calcular $t_{(2)i+1}$ y $t_{(1)i+1}$. Las iteraciones pueden continuar hasta un cierto criterio de paro (hasta obtener todos los componentes PLS que se desean) o hasta alcanzar la dimensión total de X (P_1). Tanto el algoritmo como el método de potencia convergen rápidamente (lo cual no ocurre cuando los valores propios son igual de grandes en X y Y), es decir, típicamente en menos de 10 iteraciones. En el método de potencia la convergencia ocurre de mediante las siguientes ecuaciones:

$$\begin{aligned} YY^TXX^Tt_{(2)} &= at_{(2)} && (1.1)\\ Y^TXX^TYc &= ac && (1.2)\\ XX^TYY^Tt_{(1)} &= at_{(1)} && (1.3)\\ X^TYY^TXw &= aw. && (1.4) \end{aligned}$$

Nótese en las ecuaciones anteriores que a es el valor propio más grande, y que maximiza las formas cuadráticas anteriores. Los vectores $t_{(2)}$, c, $t_{(1)}$ y w son los vectores propios asociados al valor propio a en cada una de las matrices. Esto podría ser un argumento fuerte, por lo que a continuación se presentan algunas interpretaciones con el objetivo de que se tenga un mayor entendimiento de la idea anterior.

Interpretaciones

Algo mencionado pocas veces en los artículos de PLS es la geometría y las propiedades del algoritmo de regresión PLS. Por ejemplo, supóngase que la matriz X se multiplica por una matriz ortogonal O_x, es decir, la matriz X se rota en alguna dirección, a este producto se denotará por S. Recordando que $tr(X^TX)$ es la suma de los valores propios de X^TX, entonces $tr(X^TX)$ es el total de variación, más aún $tr(X^TX) = tr(S^TS)$, por lo tanto la variación total es invariante bajo rotaciones. Si se aplica el mismo análisis a Y con la matriz O_y, entonces se tiene la matriz Z la cual es una rotación de la matriz Y. La pregunta clave en este caso es ¿Qué tan cerca están las columnas de S de las columnas de Z? considerando que $N > P_1$. El problema se puede plantear de la siguiente forma

$$\min_{o_x o_y} \sum_{i=1}^{p_1} |s_i - z_i|^2 + \sum_{i=p_1+1}^{n} |s_i|^2 \qquad (1.5)$$

La idea es la siguiente: se rota el espacio generado por la matriz X con respecto a la matriz O_x, la cual le proporciona una orientación diferente al espacio original, de forma paralela se realiza el mismo procedimiento para la matriz Y con la matriz O_y. Las matrices O_x y O_y se escogen de tal forma que ambos espacios generados sean lo más similar posible. Esto es con el objetivo de que al proyectar el espacio rotado de la matriz Y sobre el espacio rotado de la matriz X la correlación aumente y se obtenga un modelo robusto. Por lo tanto la idea es rotar los espacios conformados por las matrices X y Y, de tal forma que los espacios estén

lo más cerca posible el uno del otro, es decir, se crea un juego de espacios entre X y Y, para maximizar las correlaciones entre esos espacios. Con ésta idea se pretende que se tenga una visión más clara de la posible interpretación de las ecuaciones [(1.1)-(1.4)]. Haciendo otras analogías en términos matemáticos se obtienen los siguientes puntos

- Interpretación 1: La ecuación (1.5) puede ser escrita como

$$tr(X^TX) + tr(Y^TY) - 2tr(X^TYO_yO_x^T)$$

  Por lo tanto, si se quiere minimizar la ecuación anterior basta con maximizar el último término, el cual se puede ver como algo muy parecido a la covarianza entre las matrices rotadas X y Y.

- Interpretación 2: Los vectores w y c en el algoritmo PLS satisfacen la maximización

$$\text{máx}[Cov(s,z)]^2 = [Cov(XO_x, YO_y)]^2 = [Cov(t_{(1)}, t_{(2)})]^2$$

  Por lo tanto, los componentes $t_{(1)}$ y $t_{(2)}$ tienen la interpretación de que estos son los componentes en el espacio de X y Y que tienen la máxima covarianza entre todos los componentes de los espacios de X y de Y. El algoritmo sólo selecciona una pareja de componentes a la vez, porque la covarianza del segundo par es más chica que la covarianza más grande de la siguiente iteración.

- Interpretación 3: PLS como regresión en componentes ortogonales

  En una regresión lineal las variables son seleccionadas en base a la covarianza de la matriz X^TX y posteriormente se hace la regresión con dichas variables. Pero en lugar de realizar el proceso anterior, primero se identifica una matriz Γ de tal forma que rota a X, y luego se procede con la regresión de Y sobre el espacio generado por $X\Gamma$, de esta forma se tendría una matriz de covarianzas ponderadas con respecto a $V = \Gamma^T\Gamma$. Para el análisis anterior la pregunta es ¿Que tan real o sensato sería usar $\Gamma = Y$?, la respuesta sería muy factible en el sentido que se esta ponderando a la matriz X con respecto a la matriz Y, por lo tanto para valores pequeños en las columnas de Y el peso será menor en las columnas de X.

Se puede decir que el argumento central de PLS radica en la construcción de ecuaciones de regresión, ya que los componentes son calculados en base a éstas; de igual forma se puede encontrar la expresión para obtener el objetivo inicial, que es relacionar a las X con las variables Y. En adición a lo anterior PLS usa ecuaciones con muchos parámetros para construir los componentes $\mathbf{t}$, esto lo hace flexible pero sensible a errores aleatorios en la información.

Con la interpretación 3, se ve que el método de PLS tiene similaridades con regresión en componentes principales, pero la diferencia radica en que PLS realiza constructos o variables latentes tanto de las X como de las Y. PLS es considerado útil cuando tienes matrices

con pocas observaciones y muchas variables. Por lo que se puede inferir que la intención de PLS es crear componentes $\mathbf{t}$ que capturen la mayor información de las X con el objetivo de predecir a Y, al mismo tiempo que se reduce la dimensionalidad de las variable predictoras. Visto de una forma más simple, los componentes $\mathbf{t}$ son combinaciones de la matriz X que mejor predicen a la combinación lineales de la matriz Y con mayor información, a través de maximizar la estructura de ecuaciones de regresión que es equivalente a maximizar las covarianzas entre las combinaciones lineales de X y Y.

A continuación se presentan las propiedades más conocidas para cuando se trabaja con componentes PLS

Propiedades

- Propiedad 1: Los vectores w_i son mutuamente ortogonales, es decir, se tiene que

$$w_j^T w_i = 0 \qquad \text{para } i \neq j.$$

- Propiedad 2: Los vectores t son mutuamente ortogonales, es decir, se tiene que

$$t_j^T t_i = 0 \qquad \text{para } i \neq j.$$

- Propiedad 3: Los vectores w_i son ortogonales a los vectores p_j para $i < j$, es decir, se tiene que

$$w_i^T p_j = 0 \qquad \text{para } i < j.$$

- Propiedad 4: Los vectores p_i son ortogonales en el espacio kernel (espacio columna) de X, es decir, se tiene que

$$p_i^T (X^T X)^{-1} p_j = 0 \qquad \text{para } i \neq j.$$

Estas cuatro propiedades se siguen de la forma en que están construidas las matrices X_i, las cuales son matrices de residuales de la regresión del paso anterior. Las demostraciones de estas propiedades se muestran en el artículo de Höskuldsson [7].

1.2. Otro enfoque de regresión PLS

En general PLSR involucra dos conjuntos de información, un conjunto $X = [x_1, ..., x_{P_1}]$ de variables predictoras y un conjunto $Y = [y_1, ..., y_{P_2}]$ de variables respuestas. El método PLSR extrae un conjunto de componentes ortogonales en el espacio predictor, con el propósito de explicar los predictores y predecir las variables respuesta.

En este trabajo el algoritmo que se va a usar para calcular los componentes PLS se llama NIPALS (por sus siglas en inglés para Nonlinear estimation by Iterative PArtial Least Squares), el cual es básicamente la misma de idea de H. Wold. En el enfoque de NIPALS el peso de la variable predictora X es calculado de tal forma que se maximiza la correlación al cuadrado con la combinación lineal de $\mathbf{t}$ para todas las variables. El procedimiento es repetido hasta obtener una convergencia de los puntajes *proxys* de Y $(t_{2(1)})$, los pesos de Y $(w_{2(1)})$, los valores de X $(t_{1(1)})$ y los pesos de X $(w_{1(1)})$, éstos valores son secuencialmente calculados cada uno siendo función de las variables *proxys* previas. La secuencia de los *proxys* $w_{1(1)}^{(s)}$ y $w_{2(1)}^{(s)}$ obtenidas de la iteración s, converge a los vectores propios dominantes de las matrices $\frac{1}{N^2}Y^TXX^TY$ y $\frac{1}{N^2}X^TYX^TY$, respectivamente. Por lo tanto si la convergencia es alcanzada, entonces los vectores $w_{1(1)}$ y $w_{2(1)}$ satisfacen el criterio

$$\underset{||w_{1(1)}||=||w_{2(1)}||=1}{\arg\max} \; cov^2(Xw_{1(1)}, Yw_{2(1)}).$$

Los vectores propios dominantes de $\frac{1}{N^2}Y^TXX^TY$ y $\frac{1}{N^2}X^TYX^TY$ coinciden respectivamente con los vectores propios dominantes por la izquierda y por la derecha de la matriz $\frac{1}{N}Y^TX$. Por lo tanto los vectores $w_{1(1)}$ y $w_{2(1)}$ también satisfacen el criterio

$$\underset{||w_{1(1)}||=||w_{2(1)}||=1}{\arg\max} \; cov(Xw_{1(1)}, Yw_{2(1)}).$$

El análisis de covarianza cruzada es común en los enfoques de PLS con dos bloques, tales como el análisis canónico PLS entre otros. Sin embargo estos enfoques difieren en la construcción de los componentes sucesivos, ya que se obtienen mediante diferentes tipos de sustracción de residuos, en el sentido de que se pretende que disminuya el residuo siguiente lo más posible. La forma de calcular los componentes PLS con el algoritmo NIPALS equivale al algoritmo en la Figura 1.1.

En PLSR, tanto X como Y son "regresadas" en función de $t_{1(1)}$ (como se muestra en detalle en los pasos **II-V** del algoritmo de la Figura 1.1), con la finalidad de que el segundo componente explique la porción de la variabilidad que el primer componente no pudo explicar y así sucesivamente con los demás componentes restantes.

Existen varios criterios para definir un número razonable de componentes a considerar en el modelo. En este trabajo se utiliza la validación cruzada para definir el número de componentes. Una vez que se realiza lo anterior se procede a usar una regresión estándar, con la finalidad de predecir las variables respuesta a través de las variables latentes.

Algorithm PLSR

Input: $\mathbf{E_0} = \mathbf{X}$, $\mathbf{F_0} = \mathbf{Y}$. Output: $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{T_1}$, $\mathbf{T_2}$, $\mathbf{P}$, $\mathbf{Q}$.

for all $h = 1, ..., H$ do

Step 0: Initialize $t_{2(1)} = t_{2(1)}^{(0)}$

Step 1: Repeat

Step 1.1: $w_{1(h)}^{(s)} = \dfrac{E_{(h-1)}^{'} t_{2(h)}^{(s)}}{||E_{(h-1)}^{'} t_{2(h)}^{(s)}||}$

Step 1.2: $t_{1(h)}^{(s)} = E_{(h-1)} w_{1(h)}^{(s)}$

Step 1.3: $w_{2(h)}^{(s)} = \dfrac{F_{(h-1)}^{'} t_{1(h)}^{(s)}}{||F_{(h-1)}^{'} t_{1(h)}^{(s)}||}$

Step 1.4: $t_{2(h)}^{(s+1)} = F_{(h-1)} w_{2(h)}^{(s)}$

until convergence of $w_{1(h)}$.

Step 2: $p_{(h)} = \dfrac{E_{(h-1)}^{'} t_{1(h)}}{t_{1(h)}^{'} t_{1(h)}}$

Step 3: $b_{(t_{2(h)}|t_{1(h)})} = \dfrac{t_{2(h)}^{'} t_{1(h)}}{t_{1(h)}^{'} t_{1(h)}}$

Step 4: $E_{(h)} = E_{(h-1)} - t_{1(h)} p_{(h)}^{'}$

Step 5: $F_{(h)} = F_{(h-1)} - b_{(t_{2(h)}|t_{1(h)})} t_{1(h)} w_{2(h)}^{'}$

end for

Figura 1.1: Algoritmo PLSR con el enfoque de NIPALS [14].

El algoritmo que se ilustra en la Figura 1.1 supone modelos lineales, en el sentido de que se realizan regresiones lineales. Nótese primero que con la ecuación del paso 2, se tiene el modelo de regresión común, pero no con los supuestos usuales en los errores ε. Luego véase que para $h = 1$ el modelo ocurre lo siguiente

$$\begin{aligned} E_0 &= p_{(1)} t_{1(1)} + \varepsilon, \\ X_0 &= p_{(1)} E_0 w_{1(1)} + \varepsilon, \end{aligned}$$

con lo que se confirma que realmente se está explicando a las variables predictoras con los componentes PLS y por esto se obtiene la fórmula del paso 4, para obtener los residuos. El coeficiente $p_{(1)}$ es la primera columna de la matriz $\mathbf{P}$ en la primera iteración del algoritmo. De forma análoga, nótese que de la ecuación del paso 3 se tiene el siguiente modelo de regresión común

$$\begin{aligned} t_{2(1)} &= b_{(t_{2(1)}|t_{1(1)})} t_{1(1)} + \varepsilon \\ F_0 w_{2(1)} &= b_{(t_{2(1)}|t_{1(1)})} t_{1(1)} + \varepsilon \\ F_0 &= b_{(t_{2(1)}|t_{1(1)})} t_{1(1)} w_{2(1)}^T + \varepsilon \end{aligned}$$

por lo tanto se corrobora que $t_{2(1)}$ se proyecta en el espacio columna de $t_{1(1)}$ o equivalentemente al hecho de que F_0 es proyectado en el espacio columna de $t_{1(1)} w_{2(1)}^T$ y los coeficientes de dicha proyección se almacenan en la matriz $\mathbf{Q}$ en cada iteración.

PLS puede llevarse a cabo con diferentes métricas en la variable respuesta, pero la metodología original es con la variable Y continua. Algunas modificaciones pueden llevarse a cabo cuando este supuesto no se cumple. De igual forma las escalas en las variables que forman los constructos (las variables predictoras) típicamente se consideran continuas, en caso contrario debe llevarse a cabo un procedimiento para tomar en cuenta escalas más débiles (variables ordinales y nominales). De forma breve, en el marco usual de PLS nótese que cada variable predictora o de respuesta se toman en una escala medible, esto quiere decir que tiene una métrica continua, por otro lado las relaciones entre las variables y los constructos latentes son lineales.

1.3. Coeficientes de regresión en PLS

Con la notación de la sección anterior, una vez que se realiza el algoritmo NIPALS para obtener los H componentes PLS que se desean, entonces se obtiene la matriz $\mathbf{t} = (t_{1(1)}, ..., t_{1(h)}, ..., t_{1(H)})$, la cual contiene en sus columnas los componentes PLS. Ya con lo anterior se realiza una regresión por mínimos cuadrados para cada variable respuesta (en el caso multivariado), es decir, se proyecta cada variable respuesta en el espacio columna de $\mathbf{t}$ ($\zeta(\mathbf{t})$), por lo tanto se realizan P_2 regresiones. En otras palabras el modelo es el siguiente

$$y_j = \beta_{j,0} + \beta_{j,1} t_{1(1)} + \beta_{j,2} t_{1(2)} + ... + \beta_{j,H} t_{1(H)} + \varepsilon_j \qquad \forall j \in \{1, ..., P_2\}. \tag{1.6}$$

Si se denota a $\vec{\beta_j}$ como el vector con entradas $\beta_{j,i}$ para $i \in \{1, ..., H\}$, entonces el modelo sería $y'_j = \vec{\beta_j}\mathbf{t}$, donde y'_j es el vector y_j centrado. El intercepto o el término $\beta_{j,0}$ en (1.6) desaparece, ya que éste representa la media de y_j. Otro punto importante a notar es que se establece que la regresión en (1.6) se realiza por mínimos cuadrados, es decir, no se está haciendo ningún supuesto distribucional sobre los errores ε_j.

La regresión en el modelo (1.6) se interpreta en términos de la matriz $\mathbf{t}$ de las variables latentes, lo cual en ocasiones es razonable cuando $t_{1(h)}$ tiene un significado real. Cabe recordar que las variables latentes son constructos artificiales y en ciertas circunstancias no tienen un significado real, o específico en base al problema inicial. Por lo tanto lo ideal es conseguir la

forma usual de regresión $Y = X\beta^*$, lo cual se presenta a continuación partiendo del modelo $Y = \mathbf{t}\beta$ de la ecuación (1.6), primero nótese que ocurre con las matrices de residuos E_h para $h = 1$

$$\begin{aligned} E_1 &= E_0 - t_{1(1)}p_{(1)}^T \\ &= E_0 - E_0 w_{1(1)}p_{(1)}^T \\ &= E_0(\mathbf{I} - w_{1(1)}p_{(1)}^T) \end{aligned}$$

luego para $h = 2$

$$\begin{aligned} E_2 &= E_1 - t_{1(2)}p_{(2)}^T \\ &= E_1 - E_1 w_{1(2)}p_{(2)}^T \\ &= E_1(\mathbf{I} - w_{1(2)}p_{(2)}^T) \\ &= E_0(\mathbf{I} - w_{1(1)}p_{(1)}^T)(\mathbf{I} - w_{1(2)}p_{(2)}^T) \\ &= E_0 \prod_{h=1}^{2}(\mathbf{I} - w_{1(h)}p_{(h)}^T) \end{aligned}$$

continuando con la misma idea, para $h = H - 1$ y multiplicando ambos lados por $w_{1(H)}$ se tiene que

$$E_{H-1}w_{1(H)} = E_0 \left[\prod_{h=1}^{H-1}(\mathbf{I} - w_{1(h)}p_{(h)}^T)\right] w_{1(H)}. \tag{1.7}$$

Si $t_{1(h)} = E_{h-1}w_{1(h)}$ y $E_0 = X$, entonces se tiene una relación entre los componentes PLS y la matriz de predictores X. Por lo tanto basta con que se obtenga una relación entre Y y X a través de los componentes PLS. Usando (1.7) en (1.6) nótese lo siguiente

$$\begin{aligned} y_j &= \beta_{j,0} + \beta_{j,1}t_{1(1)} + \beta_{j,2}t_{1(2)} + ... + \beta_{j,H}t_{1(H)} \\ y_j - \beta_{j,0} &= \beta_{j,1}E_0 w_{1(1)} + \beta_{j,2}E_1 w_{1(2)} + ... + \beta_{j,H}E_{H-1}w_{1(H)} \\ y_j - \beta_{j,0} &= \beta_{j,1}E_0 w_{1(1)} + \beta_{j,2}E_0(\mathbf{I} - w_{1(1)}p_{(1)}^T)w_{1(2)} + ... + \beta_{j,H}E_0 \left[\prod_{h=1}^{H-1}(\mathbf{I} - w_{1(h)}p_{(h)}^T)\right] w_{1(H)} \\ y_j - \beta_{j,0} &= E_0 \underbrace{\sum_{k=1}^{H} \beta_{j,k} \left[\prod_{i=1}^{k-1}(\mathbf{I} - w_{1(i)}p_{(i)}^T)\right] w_{1(k)}}_{\beta_j^*} \qquad (1.8) \\ y_j' &= X\beta_j^*. \end{aligned}$$

Mediante la ecuación (1.8), se reescribió el modelo en la forma usual, esto es, se tiene a la variable respuesta y_j en función de las variables predictoras X, con $j \in \{1, ..., P_2\}$. Con el procedimiento anterior se consigue la ventaja de interpretar directamente la relación entre las variables originales X y Y.

1.4. Otras escalas de PLS

El enfoque de la regresión en mínimos cuadrados parciales o PLSR por sus siglas en inglés (Partial Least Squares Regression), se ha convertido en una herramienta común en muchas áreas de las ciencias económicas y sociales. PLSR tiene el concepto de usar variables cuantitativas, sin embargo, las investigaciones en el área socioeconómica generalmente usan variables cualitativas, esto se debe a que a menudo están interesadas en investigar la estructura de dependencia de un conjunto de variables respuesta en un conjunto de variables predictoras, las cuales son medidas en diferentes niveles de escala. Los niveles de escala pueden ser del tipo nominal, ordinal o en intervalos.

Un enfoque simple y tradicional para hacer frente al problema de la cuantificación de las variables predictoras cualitativas, es el de reemplazar cada predictor no cuantitativo con la matriz indicadora o la matriz *dummy* correspondiente. Dicho enfoque se puede utilizar fácilmente en cualquier contexto del análisis de regresión. Sin embargo, este enfoque no toma en cuenta el concepto de variable categórica como una sola, la cual tiene distintas categorías de respuesta, porque las categorías se analizan como variables distintas, es decir, está o no está en el nivel k de la variable categórica. En el caso donde se tienen variables ordinales la matriz indicadora hace que se pierda el orden de la escala original. El usar una matriz indicadora con la escala nóminal u ordinal, ocasiona que se aumente la dimensionalidad del problema. En el capítulo siguiente se muestra una forma de sobre pasar los incovenientes y abordar el problema con un enfoque diferente al clásico.

Capítulo 2

Mínimos cuadrados parciales no métricos

Con el fin de superar los problemas del capítulo anterior, una mejor estrategia parece ser la cuantificación de cada categoría con un valor numérico, de tal manera que cada variable cualitativa se transforma en una variable cuantitativa respectivamente para poder ser usada con la herramienta PLSR. Básicamente PLSR es un modelo de regresión muy flexible, capaz de manejar grandes conjuntos de datos con independencia entre ellos. Tanto la presencia de un número limitado de datos faltantes como la presencia de multicolinealidad y colinealidad exacta, no son obstáculos para este enfoque, de hecho puede trabajarse explícitamente el caso de más variables que observaciones [15]. Primero se verá el caso continuo, así como los algoritmos numéricos para obtener las variables predictoras con las que se realizará la regresión.

El método de PLS abarca un conjunto de algoritmos para calcular los componentes PLS; estos algoritmos consisten en varias extensiones, una de ellas son los algoritmos NIPALS por sus siglas en inglés de mínimos cuadrados parciales iterados no lineales.

Las técnicas de PLS son originalmente ideadas para el manejo de conjuntos de información que pueden representarse en espacios métricos, donde las variables impuestas en el análisis son observadas a lo largo de un intervalo o en una escala proporcional. Desafortunadamente, en muchos casos los investigadores están interesados en analizar conjuntos de datos recolectados en una escala no métrica, es decir, variables ordinales o nominales.

Dado un conjunto de variables, el algoritmo NIPALS maximiza la varianza de las variables ponderadas, lo que sería equivalente a decir que, dado dos conjuntos de variables, el algoritmo PLSR maximiza la covarianza de un componente con respecto al otro conjunto. Una pregunta común sería ¿por qué usar PLS no métrico? La respuesta es por dos inconvenientes con la información ordinal: El primero es para evitar la variabilidad en las estimaciones, lo cual se puede confrontar con el criterio basado en covarianzas descrito más adelante ya que se obtienen estimaciones robustas. El segundo inconveniente es que con PLS no se tienen

interpretaciones claras de los resultados en base a la escala que se maneja.

2.1. El enfoque no métrico

PLS se originó inicialmente como una técnica para datos o espacios continuos, a través de los años en casos donde la información es categórica u ordinal existen metodologías alternas. Entre estas metodologías no existe una que predomine sobre las otras, ya que se basan en diferentes enfoques de acuerdo al problema que se presente y esto dificulta su comparación. Dentro de las metodologías recientes y destacadas se encontraron 2 enfoques diferentes; el enfoque de "*path analysis*" [4] y el enfoque "*no métrico de PLS*" [14]. El primero fue desarrollado por Gabriele Cantaluppi a finales del año 2012 y este posee un enfoque en términos de variables latentes y ecuaciones estructurales, el cual no se abordará. El segundo fue desarrollado por Giorgio Russolillo en su tesis doctoral a mediados del año 2012, este enfoque es el que se empleará debido a que es *ad hoc* a las características del problema. Con ayuda del autor de este enfoque se desarrolló la metodología así como su implementación para transformar variables no métricas.

El enfoque no métrico para el manejo de medidas heterogéneas en términos del PLS está basado en el concepto de un escalamiento óptimo (Optimal Scaling) [14]. El escalamiento óptimo ha sido extensamente implementado en el análisis multivariado por algoritmos iterativos que pertenecen a la familia de mínimos cuadrados alternativos (ALS). La idea del escalamiento óptimo hace referencia al uso de funciones de escalamiento óptimas para variables sin métrica, de tal forma que se transformen en variables numéricas, es decir, variables con métrica. Dicho de otras palabras, se pretende brindar una etiqueta sin significado pero con métrica, de manera que se pueda trabajar con ella en el marco de PLS, éste procedimiento se denomina cuantificación. El principio del escalamiento óptimo toma las observaciones como categóricas, y representa a cada observación categórica por un parámetro de escalamiento. El parámetro de escalamiento está sujeto a restricciones que se derivan de las características de medición de las variables originales. En el enfoque ALS los parámetros están divididos en 2 subconjuntos: los parámetros del modelo y los parámetros de escalamiento, luego una función de pérdida se optimiza mediante optimización alternada, es decir, se optimiza con respecto a un subconjunto manteniendo el otro subconjunto fijo y viceversa.

El algoritmo PLS no métrico aprovecha el tipo de iteraciones NIPALS para implementar un procedimiento de escalamiento óptimo, lo cual lleva a una nueva clase de algoritmos PLS que manejan variables no medibles. Estos métodos son llamados PLS no métricos (NMPLS), esto es debido a que son capaces de proveer información con una nueva estructura métrica, cuando las variables carecen de una. La nueva estructura métrica no depende de las características de la información original, en otras palabras, los métodos NMPLS manejan los datos no métricos y proporcionan datos con una nueva métrica, haciendo las relaciones continuas entre las variables latentes y los constructos, como se requiere en los métodos de PLS estándar.

En la descripción del algoritmo de ahora en adelante se trabajará con variables centradas. Sea x^* una variable la cual ha sido medida para las N unidades, dado una escala de medición y que tiene que ser provista de una métrica. En el proceso de escalamiento óptimo un valor de escala es asignado a cada categoría ϕ_k de x^* con $k \in \{1, ..., K\}$ y $K \leq N$, de tal manera que

- ϕ_k es coherente con el nivel de escalamiento elegido;
- ϕ_k optimiza el criterio del modelo.

Cada variable renglón es transformada como $\hat{x} \propto \tilde{X}\phi$, donde $\phi^T = (\phi_1, \phi_2, ..., \phi_K)$ es el vector óptimo de parámetros de escalamiento y la matriz $\tilde{X}$ define un espacio donde se respetan las restricciones impuestas por el respectivo nivel de escalamiento. Como se mencionó anteriormente, el símbolo $\propto$ denota que el lado izquierdo de la ecuación corresponde al lado derecho normalizado con varianza unitaria.

Con el propósito de optimizar el criterio del modelo NMPLS, para cualquier variable x^* en la matriz de predictores, el vector de escalamiento correspondiente debe satisfacer el siguiente criterio:

$$\arg\max_{\phi} cor^2(\tilde{X}\phi, \gamma_{x^*}). \tag{2.1}$$

El criterio (2.1) se optimiza mediante las medias de los coeficientes OLS del criterio

$$\arg\max_{\forall w_q} \sum_q cov(X_q w_q, \gamma_q),$$

es decir, se proyecta a γ_{x^*} en el espacio definido por las columnas de $\tilde{X}$. La proyección resultante, normalizada y con varianza unitaria, es la representación geométrica de la variable escalada $\hat{x}$.

En general se pueden considerar tres niveles de escalamiento: nominal, polinomial y ordinal. Cada nivel de escalamiento tienen una función de escalamiento $\mathbf{Q}$ correspondiente, la cual es el operador de la proyección en el constructo latente en un adecuado espacio proyectado por las columnas de X. En la escala nominal implica la cuantificación de números, es decir, etiquetas numéricas con ningún significado cuantitativo, a diferencia de la escala ordinal en la cual la escala tiene in cierto orden; mientras que en la escala polinomial se dirige exclusivamente a la no linealidad, ya que implica la transformación de una variable métrica.

El escalamiento ordinal puede incluir variables ordinales o métricas, ya que conserva la propiedad de orden de x^*. Por lo tanto si X^* es una variable ordinal, se debe de buscar

cuantificaciones en un subespacio en particular. En la escala ordinal la proyección de γ_{x^*} que se utiliza es la siguiente

$$\mathbf{Q}(\tilde{X}^o, \gamma_{x^*}) = \tilde{X}^o(\tilde{X}^{o^T}\tilde{X}^o)^{-1}\tilde{X}^{o^T}\gamma_{x^*}. \tag{2.2}$$

donde X^o ("o" es de ordinal) está construido de acuerdo a la transformación monótona de mínimos cuadrados secundaria de Kruskal [10] sobre x^*. En el marco de NMPLS, el algoritmo *up-and-down block* de Kruskal se implementó para obtener $\tilde{X}^o$. Este algoritmo consiste en un conjunto de regresiones de γ_{x^*} en las matrices indicadoras. Este procedimiento se repite hasta que los coeficientes de regresión respeten la condición de monotonía. La transformación de Kruskal se calcula en R (ver apéndice C), la idea del algoritmo es que para cada columna de X (variable a transformar), cuenta los diferentes valores en la columna. Se expande la columna en una matriz dummy (similar al enfoque clásico), luego se obtiene un vector con los promedios de γ_{x^*} con respecto al valor original en la matriz X (para que respeten el orden de las categorias). Al final realiza un promedio ponderado de γ_{x^*}, donde los pesos son los unos de la matriz indicadora inicial, y con el cual se realiza una regresión con respecto a γ_{x^*} para poder obtener la transformación en una escala métrica.

El vector de coeficientes de regresión $(\tilde{X}^{o^T}\tilde{X}^o)^{-1}\tilde{X}^{o^T}\gamma_{x^*}$ contiene los valores de escalamiento óptimos no normalizados, los cuales preservan el orden de las categorías originales de x^*, de acuerdo a como se requiere por la condición

$$(x_i^* \sim x_{i'}^*) \Rightarrow (\hat{x}_i = \hat{x}_{i'}) \qquad \text{y} \qquad (x_i^* \prec x_{i'}^*) \Rightarrow (\hat{x}_i \leq \hat{x}_{i'}), \tag{2.3}$$

donde el símbolo $\prec$ implica un orden empírico.

Las funciones de cuantificación nominal y ordinal proveen de forma sencilla y clara escalas interpretables, gracias al hecho de que en el enfoque de PLS el peso de una variable es una función de su correlación con el constructo latente correspondiente. Como $0 \leq cor(\gamma_{x^*}, \hat{x}) \leq 1$, esta correlación jamás será negativa. Esto implica que la relación entre una variable generada por una función de escalamiento nominal y el constructo latente puede ser interpretada en términos de la intensidad pero no en términos del signo.

El algoritmo de Kruskal implementa la regresión monótona de γ en x^*. La varianza residual de la regresión es, como una consecuencia, un índice de la desviación de la monotonía. De hecho, este índice es igual al índice de *Stress* de Kruskal. En otras palabras, la correlación entre γ y $\hat{x}$, puede ser calculada como una función del *Stress* de la siguiente manera

$$cor(\gamma_{x^*}, \hat{x}) = \begin{cases} \sqrt{1 - Stress^2_{(\gamma_{x^*}, \hat{x})}}, & \text{si } cor(\gamma_{x^*}, \hat{x}) \geq 0; \\ -\sqrt{1 - Stress^2_{(\gamma_{x^*}, \hat{x})}}, & \text{si } cor(\gamma_{x^*}, \hat{x}) < 0. \end{cases}$$

donde el Stress para una cierta configuración x^* se calcula de la siguiente manera

$$S = Stress \quad = \quad \sqrt{\frac{S^*}{T^*}} = \sqrt{\frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}}, \tag{2.4}$$

con

$$d_{ij} \quad = \quad \left[\sum_{\ell=1}^{t} (x_{i\ell} - x_{j\ell})^2 \right]^{1/2}, \tag{2.5}$$

aquí los valores de $\hat{d}_{ij}$, son aquellas distancias euclídeas de una configuración $\hat{x}$ tal que minimizan S, sujetas a la restricción de que $\hat{d}_{ij}$, tiene el mismo orden de importancia como δ_{ij}, más precisamente, la restricción es que $\hat{d}_{ij} \leq \hat{d}_{i'j'}$, siempre que $\delta_{ij} \leq \delta_{i'j'}$.

El índice de *Stress* toma cualquier valor entre 0 y 1. Este índice expresa la procedencia de la relación entre γ_{x^*} y $\hat{x}$ del supuesto de monotonicidad. En otras palabras se pretende que el *Stress* sea una medida de lo bien que la configuración coincide con los datos; más completo, se supone que las diferencias "verdaderas" resultan de cierta distorsión monótona desconocida de las distancias entre puntos de una "verdadera" configuración, y que las diferencias observadas difieren de las verdaderas diferencias sólo por fluctuación aleatoria. El *Stress* es la raíz cuadrada media de los residuos de esta hipótesis. Por definición, la configuración que mejor ajusta en un espacio $J - dimensional$, para un valor fijo de j, es la configuración que minimiza el *stress*. El signo de $cor(\gamma_{x^*}, \hat{x})$ depende del tipo de transformación monótona que es aplicada a x^*, la cual puede ser creciente o decreciente.

2.2. Regresión PLS no métrica

Sean dos bloques de variables X^* y Y^* medidas en ciertas escalas. Se denota a la variable genérica de Y^* como y_p^* (con $p = 1, ..., P_2$), mientras que la variable genérica de X^* se denota como x_p^* (con $p = 1, ..., P_1$). La regresión PLS no métrica (NM-PLSR) busca matrices de datos escalados de manera óptima con $\hat{X}$, con la columna genérica $\hat{x}_{p1} \propto \tilde{X}_{p1}\phi_{p1}$, y $\hat{Y}$, con la columna genérica $\hat{y}_{p2} \propto \tilde{Y}_{p2}\phi_{p2}$. Para la variable de orden p (p-ésima variable), la matriz $\tilde{X}_{pq}$ ($q \in \{1, 2\}$) define las restricciones gracias al nivel de escalamiento y ϕ_{p2} representa el vector con los valores escalados.

Las escalas de NM-PLSR son óptimas en el sentido de que ellas optimizan un solo componente del criterio

$$\underset{w_{p1(1)}, w_{p2(1)}, \forall\phi_{p1}, \forall\phi_{p2}}{\arg\max} \; cov^2(\hat{Y}w_{1(1)}, \hat{X}w_{2(1)}), \tag{2.6}$$

bajo las restricciones $||w_{1(1)}|| = ||w_{2(1)}|| = var(\hat{x}_{p1}) = var(\hat{y}_{p2}) = 1$. Cabe destacar que este criterio implica sólo el primer componente, también se puede aplicar a los otros componentes

PLS. El criterio (2.6) depende de dos conjuntos de parámetros. El primer grupo consiste en los parámetros del modelo, limitados a la norma unitaria; el otro conjunto contiene a los parámetros de escala, que deben respetar las restricciones debido al nivel de escalamiento elegido para cada variable y para la restricción de normalización aplicable a las variables a escala. Aquí es donde se usa el enfoque de los NIPALS en el marco de NM-PLSR, ya que primero se fijan los parámetros de escala, y el problema de optimización de NM-PLSR se transforma en

$$\underset{||w_{1(1)}||=||w_{2(1)}||=1}{\arg\max} \; cov^2(\hat{Y}w_{1(1)}, \hat{X}w_{2(1)}). \tag{2.7}$$

Los $w_{1(1)}$ y $w_{2(1)}$ óptimos son respectivamente los vectores propios dominantes por la izquierda y por la derecha de la matriz $\frac{1}{N}\hat{Y}^T\hat{X}$. Por lo tanto, con el fin de optimizar el criterio (2.7), las condiciones

$$w_{2(1)} = \frac{\hat{Y}^T\hat{X}w_{1(1)}}{||\hat{Y}^T\hat{X}w_{1(1)}||}, \tag{2.8}$$

y

$$w_{1(1)} = \frac{\hat{X}^T\hat{Y}w_{2(1)}}{||\hat{X}^T\hat{Y}w_{2(1)}||}, \tag{2.9}$$

deben ser respetadas.

Una vez encontrados estos parámetros, se fijan y se procede a calcular los otros que supusieron fijos mediante el siguiente criterio

$$\begin{aligned} \underset{\forall\phi_{p1}}{\arg\max} \quad & cov^2(\hat{Y}w_{1(1)}, \hat{X}w_{2(1)}) \\ = \; & \frac{1}{N^2} w_{2(1)}^T \hat{Y}^T\hat{X}\hat{X}^T\hat{Y}w_{2(1)} \\ = \; & \frac{1}{N^2} t_{2(1)}^T \hat{X}\hat{X}^T t_{2(1)} \\ = \; & \sum_p^{P_1} cov^2(\hat{x}_{p1}, t_{2(1)}) \\ = \; & \sum_p^{P_1} cor^2(\hat{x}_{p1}, t_{2(1)}) var(\hat{x}_{p1}) var(t_{2(1)}). \end{aligned} \tag{2.10}$$

Como $var(\hat{x}_{p1}) = 1$ y $var(t_{2(1)})$ está fija con respecto a la suma, el criterio (2.10) puede ser reescrito de la siguiente manera

$$\arg\max_{\forall\phi_{p1}} \sum_{p}^{P_1} cor^2(\tilde{X}_{p1}\phi_{p1}, t_{2(1)}). \tag{2.11}$$

El criterio (2.11) es separable con respecto a cada ϕ_{p1}, y esto puede ser considerado como una suma de los componentes de P_1, cada uno de los cuales es una función de los parámetros de escalamiento de una sola variable:

$$\forall p \in \{1, ..., P_1\} \qquad \arg\max_{\phi_{p1}} cor^2(\tilde{X}_{p1}\phi_{p1}, t_{2(1)}). \tag{2.12}$$

El criterio de (2.12) es equivalente a optimizar el criterio (2.10). Los parámetros de escalamiento pueden ser calculados independientemente como coeficientes OLS de una regresión de $t_{2(1)}$ en cada matriz $\tilde{X}_{p1}$. Por lo tanto cada predictor escalado óptimo $\hat{x}_{p1}$ puede ser calculado como

$$\hat{x}_{p1} \propto \mathbf{Q}(\tilde{X}_{p1}, t_{2(1)}).$$

Un razonamiento intuitivo se puede utilizar para encontrar las cuantificaciones óptimas para las variables de respuesta, lo que lleva a que para cada $p \in \{1, ..., P2\}$, el criterio

$$\arg\max_{\phi_{p2}} cor^2(\tilde{Y}_{p2}\phi_{p2}, t_{1(1)}), \tag{2.13}$$

el cual se satisface con

$$\hat{y}_{p2} \propto \mathbf{Q}(\tilde{Y}_{p2}, t_{1(1)}). \tag{2.14}$$

Este tratamiento intuitivo funciona, debido al hecho de que un solo componente del modelo PLSR es simétrico. El modelo de PLSR se vuelve asimétrico sólo cuando las dimensiones latentes sucesivas se calculan mediante la sustracción de residuos (deflating) de ambos predictores y las respuestas con respecto a los componentes en el espacio predictor.

Por lo tanto se sigue el algoritmo en la Figura 2.1 para obtener los componentes PLS, donde $X_1 = X$ y $X_2 = Y$.

2.3. Intervalos de confianza Bootstrap

2.3.1. Intervalos de confianza

Antes de comenzar con la terminología de Bootstrap, se dará una breve introducción sobre las ideas y la metodología de los intervalos de confianza, así como el significado de la

precisión de los intervalos de confianza. Primero, sea $\hat{\theta}$ un estimador que se distribuye normal con esperanza desconocida θ, es decir, $\hat{\theta} \sim N(\theta, se^2)$, con error estándar se conocido. Para muestras grandes ($n \to \infty$) se cumple que

$$\frac{\hat{\theta} - \theta}{\hat{se}} \sim N(0, 1). \tag{2.15}$$

La igualdad $P(|Z| \leq z^{(1-\alpha/2)}) = 1 - \alpha$ es algebraicamente equivalente a

$$P_\theta(\theta \in [\hat{\theta} - z^{(1-\alpha/2)}se, \hat{\theta} + z^{(\alpha/2)}se]) = 1 - \alpha,$$

donde $z^{(\alpha)}$ es el percentil de probabilidad α de una variable aleatoria con distribución normal estándar. Por conveniencia se denotará a los intervalos de confianza por $[\hat{\theta}_l, \hat{\theta_u}]$, entonces $\hat{\theta}_l = \hat{\theta} + z^{(\alpha/2)}se$ y $\hat{\theta}_u = \hat{\theta} - z^{(1-\alpha/2)}se$. En este caso se observa que el intervalo $[\hat{\theta}_l, \hat{\theta}_u]$ tiene probabilidad exacta a $(1 - \alpha)$ de contener el verdadero valor de θ. Un intervalo de confianza de $(1 - \alpha)$ donde $P_\theta(\theta < \hat{\theta}_l) = P_\theta(\theta > \hat{\theta}_u) = \alpha/2$ es llamado intervalo de colas iguales o intervalo simétrico.

2.3.2. Bootstrap

El *Bootstrap* es un método de simulación por remuestreo de la información ya obtenida, para la inferencia estadística. El uso del término Bootstrap proviene de la frase en inglés *pull oneself up by one's bootstrap*[5]. Cada vez que se obtiene una muestra con reemplazamiento, se dice que se tiene una *muestra bootstrap*. La repetición de las muestras bootstrap son *replicaciones bootstrap*. En otras palabras, se tiene una muestra $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, y de la muestra completa se calcula el *estadístico original* (θ) el cual depende de la muestra, entonces, se obtiene $\theta = \theta(\mathbf{x})$. Por otro lado si se realiza un muestreo con reemplazamiento de tamaño n, se consigue la primera réplica bootstrap denotada por $\mathbf{x}_1^*$, ésta réplica permite calcular el estadístico de interés, así se conseguirá el primer estadístico bootstrap $\hat{\theta}_1^* = \theta(\mathbf{x}_1^*)$. Este procedimiento se realiza R veces, de tal forma que se tenga un vector (caso univariado) o una matriz (caso multivariado) de estadísticos bootstrap $\hat{\theta}_{Boot} = (\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_R^*)$. En el marco teórico de Bootstrap la inferencia de $\hat{\theta}$ se realiza sobre $\hat{\theta}_{Boot}$.

2.3.3. Intervalos por percentiles

Los intervalos por percentiles son comúnmente usados y genéricos, ya que carecen de un supuesto distribucional. La idea general de los intervalos por percentiles es la siguiente: Sea $\mathbf{x}$ una muestra aleatoria, de la cual se puede realizar muestreo con reemplazo. Sea $\hat{G}$ la función de distribución acumulada de $\hat{\theta}^*$. El intervalo por percentil de $1 - \alpha$ está definido por los percentiles $\alpha/2$ y $1 - \alpha/2$ de la siguiente forma

$$[\hat{\theta}_l, \hat{\theta}_u] = [\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1 - \alpha/2)]. \tag{2.16}$$

Ya que por definición $\hat{G}^{-1}(\alpha)$ es igual al percentil empírico de probabilidad α, de la distribución Bootstrap.

En el capítulo siguiente implementaremos las herramientas y los modelos estadísticos necesarios para analizar la muestra piloto, adicionalmente se interpretaran los resultados y detalles con respecto a la aplicación en este caso, para poder obtener las conclusiones. Básicamente la idea general de la estructura de la tesis es dar a conocer la teoría para poder analizar la base de datos relacionada al problema y poder sacar las conclusiones tanto de la metodología como del campo de aplicación en el problema.

Más adelante en el capítulo final, mediante el software *R* [13] se realiza la implementación de la metodología de mínimos cuadrados parciales no métricos, es decir, se hará un análisis de NM-PLSR para la muestra piloto del cuestionario y se calcularán los intervalos Bootstrap por percentiles, para los coeficientes de regresión de las variables predictoras. Por lo tanto la intención es que con todo el análisis anterior se aproveche el tipo de información que se captura, el tipo de problema al que se esta presentando, y se pueda extraer la mayor información posible y que sea interpretable en términos del problema que se presenta.

Algorithm 5 NMPLSR

Input: $\mathbf{X_1^*}$, $\mathbf{X_2^*}$. Output: $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{T_1}$, $\mathbf{T_2}$, $\mathbf{P}$, $\mathbf{B}$, $\hat{\mathbf{X}}_\mathbf{1}$, $\hat{\mathbf{X}}_\mathbf{2}$.
Step 1.0: Initialize $t_{2(1)} = t_{2(1)}^{(0)}$

Step 1.1: Repeat

for all $p = 1, ..., P_1$ do

Step 1.1.1: $\hat{x}_{p1}^{(s)} \propto Q(\tilde{X}_{p1}, t_{2(1)}^{(s)})$

end for

Step 1.1.2: $w_{1(1)}^{(s)} = \hat{X}_1^{(s)'} t_{2(1)}^{(s)} / ||\hat{X}_1^{(s)'} t_{2(1)}^{(s)}||$

Step 1.1.3: $t_{1(1)}^{(s)} = \hat{X}_1^{(s)} w_{1(1)}^{(s)}$

for all $p = 1, ..., P_2$ do

Step 1.1.4: $\hat{x}_{p2}^{(s)} \propto Q(\tilde{X}_{p2}, t_{1(1)}^{(s)})$

end for

Step 1.1.5: $w_{2(1)}^{(s)} = \hat{X}_2^{(s)'} t_{1(1)}^{(s)} / ||\hat{X}_2^{(s)'} t_{1(1)}^{(s)}||$

Step 1.1.6: $t_{2(1)}^{(s+1)} = \hat{X}_2^{(s)} w_{2(1)}^{(s)}$

until convergence of $w_{1(1)}$.

Step 1.2: $p_{(1)} = \hat{X}_1^{'} t_{1(1)} / t_{1(1)}^{'} t_{1(1)}$

Step 1.3: $b_{(t_{2(1)}|t_{1(1)})} = t_{2(1)}^{'} t_{1(1)} / t_{1(1)}^{'} t_{1(1)}$

Step 1.4: $E_{(1)} = \hat{X}_1 - t_{1(1)} p_{(1)}^{'}$

Step 1.5: $F_{(1)} = \hat{X}_2 - b_{(t_{2(1)}|t_{1(1)})} t_{1(1)} w_{2(1)}^{'}$

for all $h = 2, ..., H$ do

Step 2.0: Initialize $t_{2(h)} = t_{2(h)}^{(0)}$

Step 2.1: Repeat

Step 2.1.1: $w_{1(h)}^{(s)} = E_{(h-1)}^{'} t_{2(h)}^{(s)} / ||E_{(h-1)}^{'} t_{2(h)}^{(s)}||$

Step 2.1.2: $t_{1(h)}^{(s)} = E_{(h-1)} w_{1(h)}^{(s)}$

Step 2.1.3: $w_{2(h)}^{(s)} = F_{(h-1)}^{'} t_{1(h)}^{(s)} / ||F_{(h-1)}^{'} t_{1(h)}^{(s)}||$

Step 2.1.4: $t_{2(h)}^{(s+1)} = F_{(h-1)} w_{2(h)}^{(s)}$

until convergence of $w_{1(h)}$.

Step 2.2: $p_{(h)} = E_{(h-1)}^{'} t_{1(h)} / t_{1(h)}^{'} t_{1(h)}$

Step 2.3: $b_{(t_{2(h)}|t_{1(h)})} = t_{2(h)}^{'} t_{1(h)} / t_{1(h)}^{'} t_{1(h)}$

Step 2.4: $E_{(h)} = E_{(h-1)} - t_{1(h)} p_{(h)}^{'}$

Step 2.5: $F_{(h)} = F_{(h-1)} - b_{(t_{2(h)}|t_{1(h)})} t_{1(h)} w_{2(h)}^{'}$

end for

Figura 2.1: Algoritmo NMPLSR con el enfoque de NIPALS [14].

Capítulo 3

Un ejemplo: Tabaquismo

De acuerdo a la Organización Mundial de la Salud (OMS), el consumo de tabaco y la exposición a su humo son las principales causas de muerte prevenibles en el mundo, estimándose sus consecuencias anuales alrededor de 6 millones de muertes prematuras y cientos de miles de millones de dólares en pérdidas económicas (OMS, 2011). El diccionario de la Real Academia de la lengua Española (RAE) define el tabaquismo como la intoxicación crónica producida por el abuso del tabaco. De forma más coloquial se puede decir que el tabaquismo es el síndrome de dependencia del consumo del tabaco, ya sea a través de la exposición al humo de sus hojas o a través de masticarlas.

Los problemas de salud provocados por el tabaquismo dependen de la frecuencia, del tiempo de exposición, y de la vía de administración. En el caso del cigarrillo, la exposición no es sólo a la nicotina sino a los productos de la combustión, de los cuales se han identificado miles. Entre estos productos destacan el monóxido de carbono (CO), que tiene mayor afinidad con la hemoglobina que el oxígeno; el alquitrán; diversos hidrocarburos policíclicos como el fenantreno; diversas nitrosaminas; y metales como el níquel, cadmio, cromo y arsénico. Los efectos deletéreos de la exposición pasiva al humo de cigarrillo también son mensurables, y en términos generales hay correspondencia entre el grado de exposición y las consecuencias adversas.

Algunos de los principales problemas de salud relacionados con el tabaquismo son el cáncer de pulmón, el enfisema pulmonar, la EPOC[1], entre otros. Una característica de esta enfermedad que la hace difícil de combatir es la ignorancia del nivel de riesgo individual que tienen los sujetos expuestos al tabaquismo, es decir, una persona no está consciente del nivel de riesgo que tiene de desarrollar algún tipo de complicación, sea ésta cáncer, enfisema, o algún otra.

Si bien la evidencia descarta la posibilidad de que el tabaquismo sea inocuo para un indi-

[1]La enfermedad pulmonar obstructiva crónica (EPOC), es un trastorno pulmonar que se caracteriza por la existencia de una obstrucción de las vías aéreas generalmente progresiva y no reversible.

viduo, también señala con la claridad que hay fluctuaciones en el riesgo de padecer algunas o todas las complicaciones antes mencionadas, en individuos y poblaciones específicas. Al tratar de explicar los riesgos se han invocado efectos físicos, efectos aditivos, y efectos genéticos, entre otros. Además, ninguno de los problemas de salud relacionados con el tabaquismo que están entre las primeras diez causas de muerte se manifiestan sólo en fumadores. Por lo tanto, es importante investigar cuál es la relación entre la dependencia del tabaco y las complicaciones, en caso de que exista.

3.1. Antecedentes del ejemplo

Se diseñó un cuestionario para medir las características de los fumadores activos y pasivos (Ápendice A), que pudieran estar relacionadas con las concentraciones de cuatro sustancias en los líquidos corporales, sustancias que servirán como biomarcadores de la exposición al humo del cigarrillo que se describirán más adelante. Es importante comentar que el cuestionario no se sometió a ningún tipo de validación, las preguntas son cerradas pero las respuestas posibles varían entre 2 y 8 posibles categorías y no tiene reglas de puntuación, por lo tanto solo se usó una parte del cuestionario. Si se asume como variables las respuestas a ciertos campos o preguntas específicas, algunas tienen un nivel de medición nominal, mientras que otras tienen un nivel ordinal. Dentro de las preguntas se seleccionaron las siguientes:

i) Edad.

ii) Índice de Masa Corporal o Índice de Quetelet

$$\text{IMC} = \frac{\text{peso} \quad (kg)}{\text{altura}^2 \, (m^2)}.$$

iii) Nivel educativo más alto alcanzado.

iv) Nivel de ingreso familiar mensual (MN).

v) Durante los pasados 15 días, ¿Cuántos días fumó?

vi) Durante los pasados 15 días, en los días en que fumó, ¿Cuántos cigarrillos consumió?

vii) ¿Con qué frecuencia alguna persona fuma en lugares cerrados en su presencia?

viii) ¿Con qué frecuencia alguien fuma dentro de su casa?

ix) ¿Consume carnes rojas?

x) ¿Ha estado expuesto al humo del carbón o leña?

xi) ¿Consume frutas y verduras diariamente?

XII) ¿Consume alcohol?

Para verificar bioquímicamente el grado de exposición al tabaquismo, se seleccionaron cuatro biomarcadores de acuerdo a la revisión bibliográfica previamente realizada. Estos biomarcadores son: el monóxido de carbono en la sangre, el cual se mide como una concentración continua de masa por volumen, y se espera que a niveles altos de esta sustancia estén asociados a una mayor exposición al tabaquismo; el nivel continuo de cotinina en sangre, que es un metabolito intermedio de la nicotina, es decir, un producto intermedio del proceso entre su neutralización y su expulsión en el cuerpo, el cual se piensa que que vaya de mayor a menor entre fumadores intensos y no fumadores; y los productos del metabolismo del fenantreno y la nitrosamina en orina, también se midieron en una escala continua.

Más adelante, se explora la muestra piloto que se levantó, y se busca una relación entre los reactivos del cuestionario que se aplicó, que ilumine el fenómeno del tabaquismo. Por otra parte, en el capítulo final se utilizará PLS, y bajo ciertas modificaciones aquí presentadas se realizará un modelo predictivo con respecto a los biomarcadores, además de establecer contrastes clasificatorios entre las variables, de acuerdo a nivel de exposición.

Las preguntas seleccionadas, consignadas arriba, tienen respuestas ordinales o continuas en el estudio. Se analizarán las respuestas de forma exploratoria, cuando se categorizaron las respuestas continuas, esto es, discretizar con categorías que tienen un orden explícito de acuerdo a la naturaleza de la respuesta. Por ejemplo si se desea categorizar la variable edad, se crean los intervalos $\mathbf{A_1} = [18, 21]$, $\mathbf{A_2} = [22, 25]$ y $\mathbf{A_3} = [25, \infty)$, estas categorías tienen un orden explícito, es decir, los sujetos que están en $\mathbf{A_1}$ son menores que los sujetos que están en $\mathbf{A_2}$ que a su vez son menores que los sujetos en $\mathbf{A_3}$. Cabe mencionar que no está comprobada una relación directa entre la edad y el riesgo a desarrollar enfermedades respiratorias, aunque intuitivamente pareciera ser cierto. Análogamente con el IMC, si el IMC está en el intervalo $\mathbf{IMC_1} = [16, 18.5)$, el sujeto tiene desnutrición o delgadez, a diferencia que estuviera en la obesidad o el sobrepeso, que corresponde al intervalo $\mathbf{IMC_3} = [25, \infty)$, por lo tanto en $\mathbf{IMC_2} = [18.5, 25)$ existen sujetos con peso normal y el orden de estos intervalos es ascendente. Para los demás reactivos, nótese que el nivel de ingreso familiar tiene 8 respuestas diferentes, las cuales se pueden convertir en 3 categorías, y el orden *3*, *2* y *1* de estas categorías dependerá de los valores que tome la variable. Por lo tanto, en el ingreso familiar mientras menos ganen el orden es de 1 y si ganan mucho el orden es de 3.

Sea $\mathbf{X}_1$ una matriz de $N \times P_1$, donde N representa a los sujetos que contestaron el estudio y con $P_1 = 12$ columnas que representan las preguntas que cada sujeto contestó, ya sea de forma ordinal o continua. Así en la siguiente tabla se presenta la estructura de la base de datos.

Edad	IMC	Nivel Educativo	Nivel de Ingresos	Cuantos días fumó de los 15 días pasados	Cigarrillos consumidos en 15 días	Presencia de humo	Frec. de fum. en casa	Carnes rojas	Exposición al humo	Consumo de frutas y verduras	Consumo de alcohol (sem)
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,6}$	$X_{1,7}$	$X_{1,8}$	$X_{1,9}$	$X_{1,10}$	$X_{1,11}$	$X_{1,12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_{N,1}$	$X_{N,2}$	$X_{N,3}$	$X_{N,4}$	$X_{N,5}$	$X_{N,6}$	$X_{N,7}$	$X_{N,8}$	$X_{N,9}$	$X_{N,10}$	$X_{N,11}$	$X_{N,12}$

donde $X_{i,j}$ es el orden del sujeto i en la pregunta j, con $i \in \{1, ..., N\}$ y $j \in \{1, ..., 12\}$.

3.2. La base de datos

La base de datos con la que se va a trabajar es una muestra piloto de una población mexicana, en la cual se midieron las características correspondientes en el estudio, obteniendo tanto las variables predictoras como las variables respuesta en los análisis clínicos, para los $N = 99$ individuos. Cabe aclarar que cada vez que se mencione base de datos se hace referencia a esta muestra piloto. Por lo tanto si se toman en cuenta los 12 reactivos principales se obtiene la Tabla 3.1.

Por otra parte, además de la información en esta tabla, se cuenta con información sobre el género de los sujetos, y si son fumadores activos o pasivos, de acuerdo a lo reflejado en sus respuestas a estas preguntas en la encuesta. A continuación se presenta el siguiente análisis descriptivo de la base de datos.

Tabla 3.1: Algunos sujetos en la Base de datos de la muestra.

ID	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}
1	1	3	2	1	2	1	3	2	2	3	1	3
2	1	2	2	1	1	1	1	1	2	1	1	2
3	2	3	2	1	1	2	1	1	2	2	1	3
4	3	2	2	2	1	1	1	1	2	1	1	1
5	3	3	2	2	1	1	3	1	2	2	1	3
6	3	2	1	2	3	2	3	2	2	3	1	2
7	2	2	2	2	1	1	3	1	1	1	1	1
8	3	3	1	2	3	2	3	3	2	2	1	3
9	2	3	2	2	1	1	1	1	2	3	2	1
10	2	1	2	3	1	1	1	1	2	2	1	1
11	2	3	2	1	1	2	2	1	2	2	1	2
12	1	2	2	3	1	1	2	1	2	1	3	1
13	3	3	1	1	3	3	3	3	2	2	1	3
14	1	3	3	3	1	1	3	1	2	2	1	1
15	1	3	2	1	1	1	3	1	2	1	1	1
16	2	2	2	1	1	1	3	1	2	3	1	1
17	2	2	2	2	1	1	1	1	2	2	1	2
18	3	2	2	2	1	1	1	1	2	2	1	3
19	1	2	3	2	3	3	1	1	2	3	1	1
20	1	3	2	1	1	1	2	1	2	2	1	1
21	2	3	2	2	1	2	3	3	2	3	1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	2	2	2	1	3	3	3	1	2	2	2	2
91	1	3	2	2	3	3	3	2	2	2	1	3
92	1	3	2	2	1	1	2	1	2	1	1	1
93	2	2	2	1	2	3	2	1	1	1	1	1
94	1	3	2	1	3	1	3	3	2	1	3	2
95	1	3	2	2	3	3	3	1	2	3	2	3
96	2	1	2	1	2	1	3	1	2	2	1	1
97	1	2	2	2	2	2	2	1	2	3	2	1
98	1	2	2	1	3	1	2	1	1	1	1	2
99	2	3	2	2	3	2	1	1	2	1	1	3

3.2.1. Análisis descriptivo

En la base de datos se presentan 99 sujetos. Los cuales se pueden clasificar de diferentes formas, una forma sería por el sexo al nacer de los sujetos. Para nótese que el 52.5 % son mujeres. Si se hace énfasis en las edades, la distribución de las edades en la muestra está en la Figura 3.1.

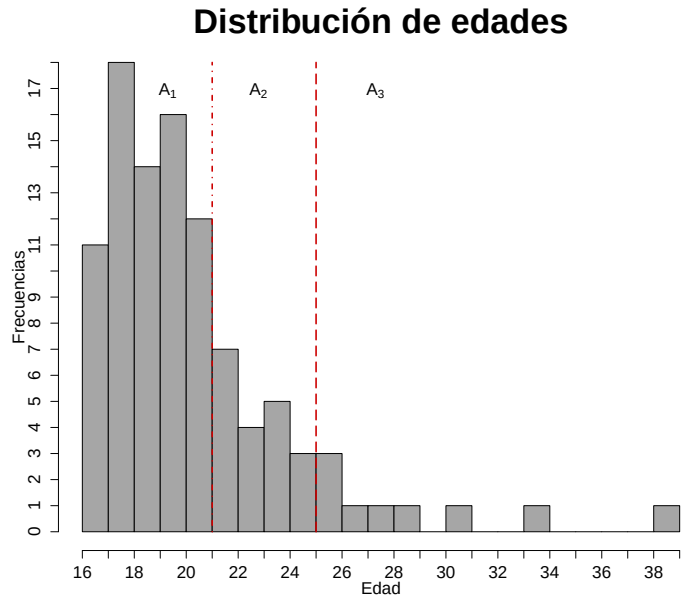

Figura 3.1: Distribución de las edades, segmentadas de acuerdo a la categorización que se hizo para analizarlos.

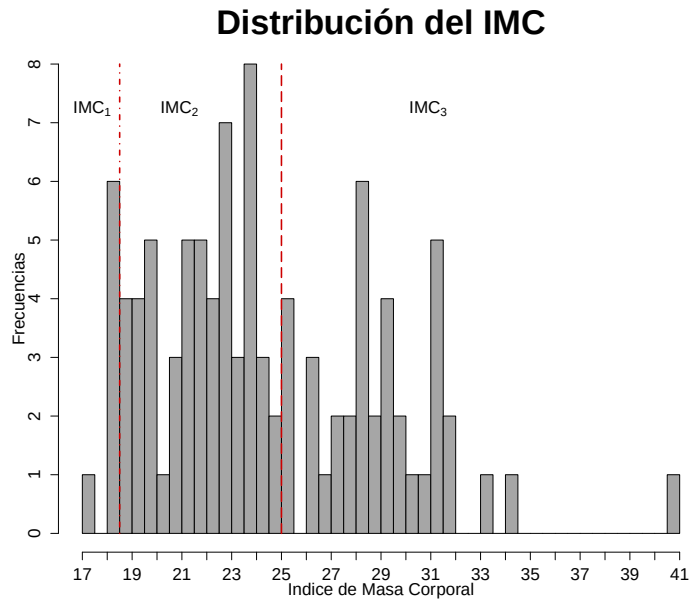

Figura 3.2: Distribución del IMC, segmentado de acuerdo a la categorización que se hizo para analizarlos.

De la Figura 3.1 se puede observar que aproximadamente el 59 % están en la categoría 1 y el 12 % en la categoría 3, es decir, aproximadamente de cada 5 personas en la base de datos 3 son relativamente jóvenes. Con respecto al IMC, la distribución en la base de datos

se refleja en la Figura 3.2 donde aproximadamente el 54 % están en la categoría 2 y 38 % en la categoría 3, con un sesgo positivo. Lo anterior indica que aproximadamente de cada 5 personas 2 tiene algún tipo de sobre peso. Parte de esto se puede explicar por los altos índices de obesidad reportados a nivel nacional. En cuanto al nivel educativo más alto alcanzado predomina la licenciatura con un 78 %, esto es debido a que el estudio se aplicó en una universidad. Por otro lado esto también repercute en el nivel económico familiar, ya que predomina una de las tres clases en la categorización.

Otro rasgo que hay que remarcar en la base de datos, es la cantidad de fumadores activos contra la de fumadores pasivos, y el número de personas que realizan actividad deportiva, contra los que no realizan alguna actividad física. Con la clasificación previa se tiene que 2/3 de los individuos son fumadores pasivos, es decir, existe el doble de fumadores pasivos que los fumadores activos y aproximadamente el 73 % de las observaciones en la base de datos tiene actividad física. Otra forma de visualizar a los sujetos en la muestra con respecto al sexo al nacer y su rol de fumador de forma conjunta es a través de la siguiente Tabla 3.2 donde se destaca que el 38 % de las mujeres en la muestra es fumadora activa, por el otro lado, el 28 % de los hombres en la muestra son fumadores activos. Visto desde otra perspectiva y si se enfoca únicamente en los fumadores activos en la muestra, entonces el 61 % son mujeres.

Tabla 3.2: Tabla de conteos para las variables clasificadoras sexo y fumador.

Fumador / Sexo	Pasivo	Activo	Total
Mujeres	32 (32.3 %)	20 (20.2 %)	52
Hombres	34 (34.3 %)	13 (13.1 %)	47
Total	66	33	99

En el cuestionario que se plantea a los sujetos, se pregunta la marca de cigarrillos que usualmente consumen los fumadores activos. En la cual existe una amplia gama de marcas de cigarrillos, en la Tabla 3.3 se muestra el consumo de las marcas más consumidas con respecto al sexo y en general. Véase como la marca Marlboro tiene al 61 % de los fumadores, mientras que Benson y Pall Mall tienen el 15 % y el 9 % respectivamente. Este fenómeno de las marcas pudiera darse debido a la región o lugar geográfico en la cual se levantó la muestra.

Hay que tener presente que esta es una muestra parcial, esto quiere decir que el estudio actualmente continúa reclutando más sujetos. Para los propósitos en la tesis, se desearía por un lado que la muestra sea hasta cierto punto heterogénea con el fin de obtener tanto mediciones de hombres como de mujeres, de fumadores activos como pasivos, con proporciones muy cercanas, pero por otro lado que sea homogénea para establecer límites o niveles de riesgo coherentes. Todo lo anterior es con la finalidad de tener conclusiones más concretas y

Tabla 3.3: Tabla de conteos por marca para la variable sexo en los fumadores activos.

Sexo \ Marca	Marlboro	Benson	PallMall	Otras	Total
Mujeres	12	4	2	2	20
Hombres	8	1	1	3	13
Total	20	5	3	5	33

así evaluar la predicción de acuerdo a ciertas clasificaciones.

3.3. Análisis de correspondencia

El análisis de correspondencia es un método que se usa para el análisis de datos categóricos. El término análisis de correspondencia se origina en Francia, donde es muy popular gracias a Jean Paul Benzecri y sus asociados, ellos formularon el análisis de correspondencia en 1960, y su trabajo culminó en la referencia Benzecri (1973) y la serie que lleva por nombre "*Practique de l'analyse des Données*". Una posible razón del retraso del uso del análisis de correspondencia fuera de Francia es atribuida a cuestiones del idioma.

3.3.1. Desarrollo

El análisis de correspondencia es una técnica con la cual es posible encontrar una representación multidimensional de la dependencia entre filas y columnas de una tabla de contingencia. El análisis de correspondencia constituye el equivalente de componentes principales para variables cualitativas (escala categórica). La representación que despliega el análisis de correspondencia es encontrada mediante la localización de los totales tanto de los renglones, como de las columnas y despliega las categorías como puntos. Estos puntos pueden ser normalizados de tal forma que las distancias entre puntos filas y/o puntos columnas en el espacio euclidiano sean iguales a las llamadas *distancias Ji-cuadradas*, las cuales se muestran más adelante.

El análisis de correspondencia puede interpretarse de dos formas equivalentes. La primera, como una manera de representar las variables en un espacio de menor dimensión, de forma análoga a componentes principales, pero definiendo la distancia entre los puntos de manera coherente con la interpretación de los datos y en lugar de utilizar la distancia euclideana se utiliza la distancia Ji-cuadrada. La segunda interpretación está más próxima al escalado multidimensional, un procedimiento cuyo objetivo es asignar valores numéricos a variables cualitativas.

3.3.2. Búsqueda de la mejor proyección

Sea X una tabla de contingencia de $I \times J$ con frecuencias no escaladas o conteos, la cual tiene elementos x_{ij}. Asúmase que X es de rango completo ($rango(X) = J$). Si n es el total global de frecuencias en X, entonces la matriz de proporciones $P = \{p_{ij}\}$ resulta de dividir cada elemento de X entre n, y por lo tanto se tiene que

$$p_{ij} = \frac{x_{ij}}{n}, \text{ con } i = 1, 2, ..., I, \text{ y } j = 1, 2, ..., J, \qquad \text{ó} \quad P = \frac{1}{n}X. \tag{3.1}$$

La matriz P es llamada matriz de frecuencias relativas o matriz de correspondencia, ahora se define r_i y c_j de la siguiente manera:

$$\begin{aligned} r_i &= \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, \qquad i = 1, 2, ..., I, \qquad \text{ó} \qquad r = P\mathbf{1}_J, \\ c_j &= \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}, \qquad j = 1, 2, ..., J, \qquad \text{ó} \qquad c = P\mathbf{1}_I, \end{aligned} \tag{3.2}$$

donde $\mathbf{1}_J$ es un vector de $J \times 1$ con el elemento uno en todas sus entradas y análogamente para $\mathbf{1}_I$ de dimensión $I \times 1$, y sean:

$$D_r = diag(r_1, ..., r_I) \qquad \text{y} \qquad D_c = diag(c_1, ..., c_J), \tag{3.3}$$

por lo tanto para propósitos de escalamiento, las raíces de las matrices son

$$D_r^{1/2} = diag(\sqrt{r_1}, ..., \sqrt{r_I}) \qquad \text{y} \qquad D_c^{1/2} = diag(\sqrt{c_1}, ..., \sqrt{c_J}). \tag{3.4}$$

El análisis de correspondencia puede ser formulado como un problema de mínimos cuadrados ponderados al seleccionar $\hat{P} = \{\hat{p}_{ij}\}$, una matriz de rango reducido, para minimizar

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{r_i c_j}(p_{ij} - \hat{p}_{ij})^2 = tr[(D_r^{-1/2}(P - \hat{P})D_c^{-1/2})(D_r^{-1/2}(P - \hat{P})D_c^{-1/2})^T] \tag{3.5}$$

de la ecuación (3.5) se ve que $(p_{ij} - \hat{p}_{ij})/\sqrt{r_i c_i}$ es la entrada (i,j) de $D_r^{-1/2}(P - \hat{P})D_c^{-1/2}$.

La solución del análisis de correspondencia puede ser obtenida de la siguiente forma: Sea P la matriz que se desea analizar y sea D_r y D_c matrices diagonales con las sumas por renglones y por columnas respectivamente (se supone que $r_i > 0$ y $c_j > 0$). Sea $E = D_r\mathbf{1}\mathbf{1}^T D_c$ una matriz de $I \times J$, con elementos $e_{ij} = r_i c_j$. Cuando se calcula la descomposición en valores singulares propios de la matriz $D_r^{-1/2}(P - E)D_c^{-1/2}$ con entradas $\frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$ se tiene que

$$D_r^{-1/2}(P-E)D_c^{-1/2} = U\Lambda V^T, \tag{3.6}$$

donde $U^TU = I = V^TV$ y $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_J)$ (los valores de los elementos en Λ están en orden descendente). Para hacer una aproximación con rango D, si se define
$D \leq \text{mín}(J-1, I-1)$, entonces

$$\underset{I\times D}{U} \quad \underset{D\times D}{\Lambda} \quad \underset{D\times J}{V^T}.$$

Los puntos renglones y puntos columna se normalizan como se muestra a continuación

$$R = D_r^{-1/2}U \qquad \text{y} \qquad C = D_c^{-1/2}V. \tag{3.7}$$

Los cuales son vectores propios ponderados. Esto es $R^TD_rR = I = C^TD_cC$ y refleja el hecho de que las sumas por renglones y por columnas de $(P-E)$ desaparece, $\mathbf{1}^TD_rR = 0 = \mathbf{1}D_cC$, es decir, los puntos renglones son no correlacionados, como los puntos por columnas, mientras que cada dimensión de estos puntos tienen media cero y varianza uno.

La relación entre puntos filas y puntos columnas está especificado por la "fórmula de transición":

$$\tilde{R} = D_r^{-1}PC \qquad \text{y} \qquad \tilde{C} = D_c^{-1}P^TR, \tag{3.8}$$

en este contexto los valores de los perfiles se usan como pesos o ponderaciones, pero si se usan como ejes coordenados a $\tilde{R}$ y $\tilde{C}$ se obtiene un espacio con las distancias ji-cuadradas. Cuando un perfil renglón es igual al perfil del renglón promedio, la ecuación (3.8) muestra que el punto renglón será el promedio ponderado de las columnas (en el origen). Cuando para alguna columna j el valor del perfil renglón p_{ij}/r_i es más grande que el promedio c_j esta columna será atraído por el punto renglón en esta dirección. Cabe resaltar que el criterio de optimización en el análisis de correspondencia no está establecido en términos de las distancias entre conjuntos, pero sí lo está en términos de las distancias dentro el conjunto, llámese

$$\tilde{R}^TD_r\tilde{R} = \Lambda R^TD_rR\Lambda = \Lambda U^TD_r^{-1/2}D_rD_r^{-1/2}U\Lambda = \Lambda^2.$$

Por lo que se debe de tener cuidado con las interpretaciones. Si se sustituye (3.7) en (3.6) se obtiene:

$$\begin{aligned} D_r^{-1/2}(P-E)D_c^{-1/2} &= U\Lambda V^T \\ P-E &= D_r^{1/2}U\Lambda V^TD_c^{1/2} \\ P &= E + D_r^{1/2}D_r^{1/2}D_r^{-1/2}U\Lambda V^TD_c^{-1/2}D_c^{1/2}D_c^{1/2} \\ &= D_r(\mathbf{1}\mathbf{1}^T)D_c + D_rR\Lambda C^TD_c \\ &= D_r(\mathbf{1}\mathbf{1}^T + R\Lambda C^T)D_c \end{aligned} \tag{3.9}$$

La cual es conocida como la fórmula de reconstrucción.

La ecuación (3.9) muestra que análisis de correspondencia sólo tiene sentido cuando estos residuos no son simplemente un resultado de variación aleatoria. En el caso de usar o no el estadístico χ^2 de Pearson:

$$\chi^2 = n \sum_i \sum_j (p_{ij} - e_{ij})/e_{ij}. \tag{3.10}$$

En las publicaciones en francés $tr(\Lambda^2)$ es llamada la inercia total, mientras que en publicaciones en inglés χ^2/n es llamado índice del cuadrado promedio de contingencia de Pearson.

Teorema 3.3.1. *El término $r_1 c_1^T$ es una aproximación considerable $\hat{P}$, sin importar cual sea la matriz de correspondencia P de $I \times J$.*

El resultado anterior aclara que $r_1 c_1^T$ es la mejor aproximación de rango 1, y corresponde al supuesto de independencia entre las filas y columnas. Esto se corrobora con la ecuación (3.9) en la matriz E la cual coincide con el término rc^T.

Otra forma de llegar al análisis de correspondencia es mediante el análisis de la matriz P. Entonces las I filas pueden tomarse como I puntos en el espacio $\mathbb{R}^J$. Se va a buscar una representación de estos I puntos en un espacio de dimensión menor que permita apreciar sus distancias relativas. El objetivo es el mismo que con componentes principales, pero ahora se debe de tener en cuenta las peculiaridades de este tipo de datos. Estas peculiaridades provienen de que la frecuencia relativa de cada fila es distinta, lo que implica que:

1. Todos las filas (puntos en $\mathbb{R}^J$) no tienen el mismo peso, ya que algunas contienen más datos que otras. Al representar el conjunto de las filas (puntos) se debe dar más peso a aquellas filas que contienen más datos.

2. La distancia euclídea entre puntos no es una buena medida de su proximidad y se debe modificar esta distancia.

Para obtener comparaciones razonables entre estas frecuencias relativas hay que tener en cuenta la frecuencia relativa de aparición del atributo de interés. En atributos raros, pequeñas diferencias absolutas pueden ser grandes diferencias relativas, mientras que en atributos con gran frecuencia, la misma diferencia será poco importante. Una manera intuitiva de manejar las relaciones de dependencia se dan en términos de las distancias Ji-cuadradas, las cuales pueden ser calculadas tanto entre renglones como entre columnas. Para calcular las distancias entre renglones, estas distancias son calculadas como perfiles de renglones en la matriz, donde el perfil del renglón i es el vector condicional de proporciones p_{ij}/r_i. La distancia Ji-cuadrada entre dos renglones i e i' está definida por

$$\delta^2(i, i') = \sum_j \frac{(p_{ij}/r_i - p_{i'j}/r_{i'})^2}{c_j}, \tag{3.11}$$

además cuando i e i' tienen el mismo perfil $\delta^2(i,i') = 0$. La diferencia entre perfiles i e i' por columna j está divido por c_j, esto otorga menos influencia al perfil a diferencia para las columnas que tienen grandes márgenes. El concepto de la distancia Ji-cuadrada se usa en la interpretación de cierta configuración de puntos. Cuando dos puntos renglones son muy cercanos, sus perfiles deben ser similares, así mismo estos renglones están muy aproximados a las columnas, y viceversa. Cuando un punto renglón está cerca del origen, este perfil es similar a la proporción de columnas c_j. El procedimiento es análogo para las distancias entre columnas.

3.4. Análisis estadístico de la base de datos

Una vez que se cuenta con los conocimientos estadísticos relativos al análisis de correspondencia, entonces se aplicarán los análisis y técnicas a la base de datos.

3.4.1. Tabla de contingencia

De la Tabla 3.1, si se define a $x_{\ell j}$ como el número de personas que tienen el nivel ℓ en la pregunta j, entonces para calcular la tabla de contingencia de dos vías como la estructura de la Tabla 3.4, donde las columnas representan las preguntas del estudio y los renglones representan el nivel cada pregunta.

Tabla 3.4: Tabla de contingencia teórica.

Preg / Nivel	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Total
Alto	x_{A1}	x_{A2}	x_{A3}	x_{A4}	x_{A5}	x_{A6}	x_{A7}	x_{A8}	x_{A9}	x_{A10}	x_{A11}	x_{A12}	R_A
Medio	x_{M1}	x_{M2}	x_{M3}	x_{M4}	x_{M5}	x_{M6}	x_{M7}	x_{M8}	x_{M9}	x_{M10}	x_{M11}	x_{M12}	R_M
Bajo	x_{B1}	x_{B2}	x_{B3}	x_{B4}	x_{B5}	x_{B6}	x_{B7}	x_{B8}	x_{B9}	x_{B10}	x_{B11}	x_{B12}	R_B
Total	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	T

$$\text{con } C_j = \sum_{\ell \in \{A,M,B\}} x_{\ell j}, \qquad R_\ell = \sum_{j=1}^{11} x_{\ell j} \qquad \text{y} \qquad T = \sum_{\ell,\, j} x_{\ell j}.$$

Cabe mencionar que como todos los sujetos contestan las preguntas adecuadamente, entonces los totales por columna serán iguales para cualquier columna j, es decir, $C_j = C_{j'}$ para $j \neq j'$, por lo que no se tiene una tabla de contingencia en el sentido estricto de la definición, si no una tabla de conteos. La tabla de conteos obtenida, se presenta a continuación

Tabla 3.5: Tabla de conteos de la Tabla 3.1.

Orden \ Preg	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}	Q_{12}	Total
3	12	38	15	4	21	11	56	24	1	26	13	24	245
2	28	54	78	38	7	16	14	5	91	36	39	26	432
1	59	7	6	57	71	72	29	70	7	37	47	49	511
Total	99	99	99	99	99	99	99	99	99	99	99	99	1188

3.4.2. Análisis de correspondencia

Otra forma de ver el conjunto de datos de la Tabla 3.5 es de forma multivariada, es decir, tener los datos en un espacio multidimensional donde cada dimensión corresponde a una pregunta con 3 categorías posibles, de esta manera se obtiene una tabla de 12 vías con 3 niveles, por lo tanto un individuo se compone de una combinación de un vector de 12 entradas, que corresponden al nivel alto, medio o bajo en cada dimensión. Por ejemplo, supóngase que se tienen 3 preguntas Q_i, Q_j y Q_k, entonces se puede hablar de un espacio tridimensional, donde se ocupa únicamente el cuadrante positivo, de tal forma que en cada dimensión solo hay 3 valores posibles, tal y como se muestra en la Figura 3.3.

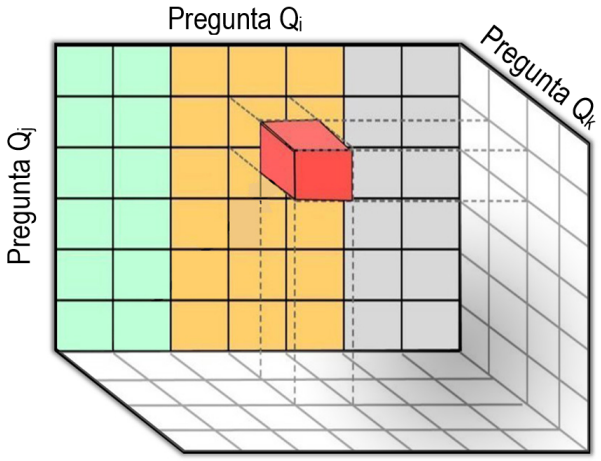

Figura 3.3: Representación gráfica de una tabla de contingencia de 3 vías.

Para el cálculo de los perfiles en los Biplots, se empleará el software SAS Base [16], con el procedimiento *proc corresp* en la paquetería $SAS/STAT$. Con el uso del lenguaje SAS, se procede a calcular las proyecciones que se necesitan para obtener el *Biplot* de la tabla de 12 vías. Los *Biplots* son representaciones gráficas en el espacio euclídeo de las proyecciones obtenidas del análisis de correspondencia multivariado[2].

[2]Las preguntas que se consideraron son las siguientes

I) Edad.

II) Índice de Masa Corporal.

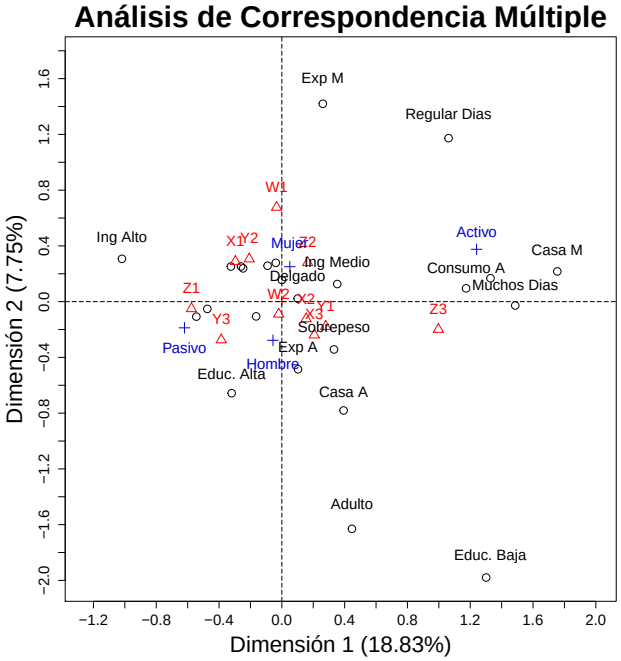

Figura 3.4: Biplot correspondiente a la tabla de 12 vías.

En la Figura 3.4 se observa que muchas de las características se agrupan en el centro, esto pudiera deberse al tamaño de muestra con el que se trabaja. Si se analizan los resultados en términos de agrupaciones se observa que en la muestra, los fumadores activos están asociados a un consumo alto, el cual mantienen un tiempo considerable de días, y así mismo también en su casa fuman en cantidades significativas. Por otra parte una posible opción es agrupar la exposición moderada con un número regular de días, lo que puede llevar a que se piense en que existe un grupo en la muestra, que tiene una exposición medianamente frecuente y esto es provocado por un consumo continuo de cigarrillos de otras personas. Con respecto al género o sexo al nacer, tanto en hombres como en mujeres ambos grupos comparten características en diferentes niveles, esto pudiera indicar que posiblemente no se pueda detectar diferencias entre hombres y mujeres con respecto al IMC, Edad, Nivel educativo, entre otras cosas. Sin embargo, en un intento de esclarecer estas conjeturas se realizó una exploración

III) Nivel educativo más alto alcanzado.

IV) Nivel de ingreso familiar mensual (MN).

V) Durante los pasados 15 días, ¿Cuántos días fumó?

VI) Durante los pasados 15 días, en los días en que fumó, ¿Cuántos cigarrillos consumió?

VII) ¿Con qué frecuencia alguna persona fuma en lugares cerrados en su presencia?

VIII) ¿Con qué frecuencia alguien fuma dentro de su casa?

IX) ¿Consume carnes rojas?

X) ¿Ha estado expuesto al humo del carbón o leña?

XI) ¿Consume frutas y verduras diariamente?

XII) ¿Consume alcohol?

más detalla de los resultados. Se detectó que por construcción, la pregunta que está relacionada con la cantidad de carne que consume el sujeto se encuentra mal planteada, es por tal motivo que se eliminará del análisis. De igual forma también se eliminarán las preguntas "¿Ha estado expuesto al humo del carbón o leña?", "¿Consume frutas y verduras diariamente?" y "¿Consume alcohol?", que corresponden a las preguntas 10, 11 y 12 descritas en el presente capítulo, que están muy próximas a los ejes en sus tres categorías, esto quiere decir que éstas 3 preguntas no están asociadas directamente a un grupo (*1*, *2* ó *3*) en específico. Cabe recordar que en los biplots del análisis de correspondencia se grafican los coeficientes resultantes de las combinaciones lineales de las proyecciones de variables ordinales, por lo tanto mientras más cerca del origen se encuentre la proyección, menor es la aportación de dicho perfil.

Una vez que se aclaró lo anterior se eliminarán las preguntas 9, 10, 11 y 12 del análisis, para después volver a realizar el mismo análisis sin esas preguntas, es así como obtenemos la Figura 3.5 con la muestra de 99 individuos, pero cabe resaltar que las relaciones pudieran cambiar cuando se tenga un muestra más grande y no se tendrían que eliminar.

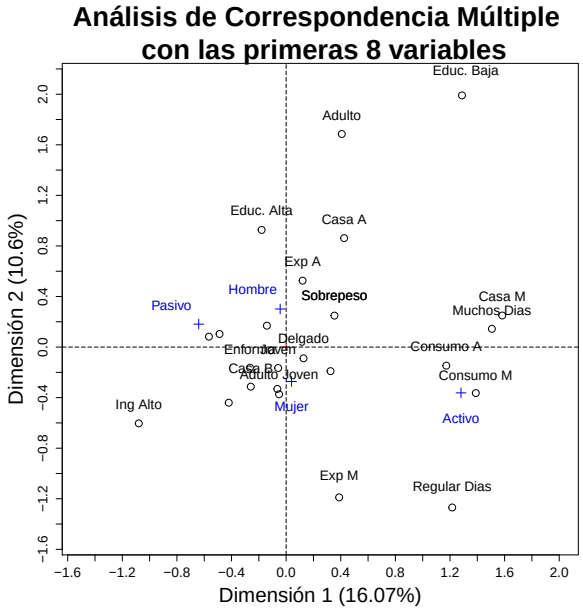

Figura 3.5: Biplot correspondientes a la Tabla de 8 vías (sin las últimas 4 preguntas).

Se puede notar que al eliminar las preguntas anteriores, entonces la dispersión en el biplot aumenta y algunas preguntas se trasladan muy poco de lugar, de forma que la relación o agrupación es la misma.

Como parte de la información que se tiene, se cuenta con las variables de sexo al nacer y si la persona es fumador activo o pasivo, es decir, se tienen variables clasificadoras por lo que algo muy intuitivo en esta situación, es realizar el análisis de correspondencia separando la

muestra de acuerdo al sexo o con respecto a su actividad como fumador pasivo o activo, de esta forma se crean escenarios y se puede extraer mayor información en cuanto a la base de datos obtenida. Si se analiza el escenario con respecto al sexo al nacer, es decir, se separan a los hombres de las mujeres y se hacen las proyecciones para cada subgrupo, entonces se obtienen los biplots en la Figura 3.6. En la Figura 3.6 se puede ver como cambia la dirección con respecto al sexo, es decir, se puede plantear el hecho de que el tabaquismo se asimila diferente con respecto al género, agrupando de forma diferentes las variables con sus niveles.

En cuanto a los hombres con un rol de fumador activo en su mayoría presentan consumos medios y altos, con un número medio de cigarrillos consumidos de forma frecuente y en su hogar mantienen una exposición media del cigarrillo, mientras que los hombres fumadores pasivos son jóvenes con un IMC normal, con niveles bajos con respecto a su consumo, su exposición fuera y dentro del hogar es de forma escasa, e ingresos altos, entre otras cosas.

Para el caso de las mujeres, las características para el grupo activo parecen estar más dispersas con respecto al de los hombres activos, dicha categoría se caracterizar con respecto a un sobrepeso, que en su casa hay una exposición alta de cigarrillo, con un consumo alto de forma regular-alta, para las mujeres con un rol pasivo en consumo del cigarrillo, se encontró una población adulto-joven, con características de bajo nivel en cuestión del consumo, de la exposición dentro y fuera de su hogar las personas no fuman, o lo hacen de forma esporádica, todo lo anterior asociado también a un ingreso alto. Sin embargo sin importar el género el análisis de correspondencia interpreta aproximadamente el mismo nivel en los datos.

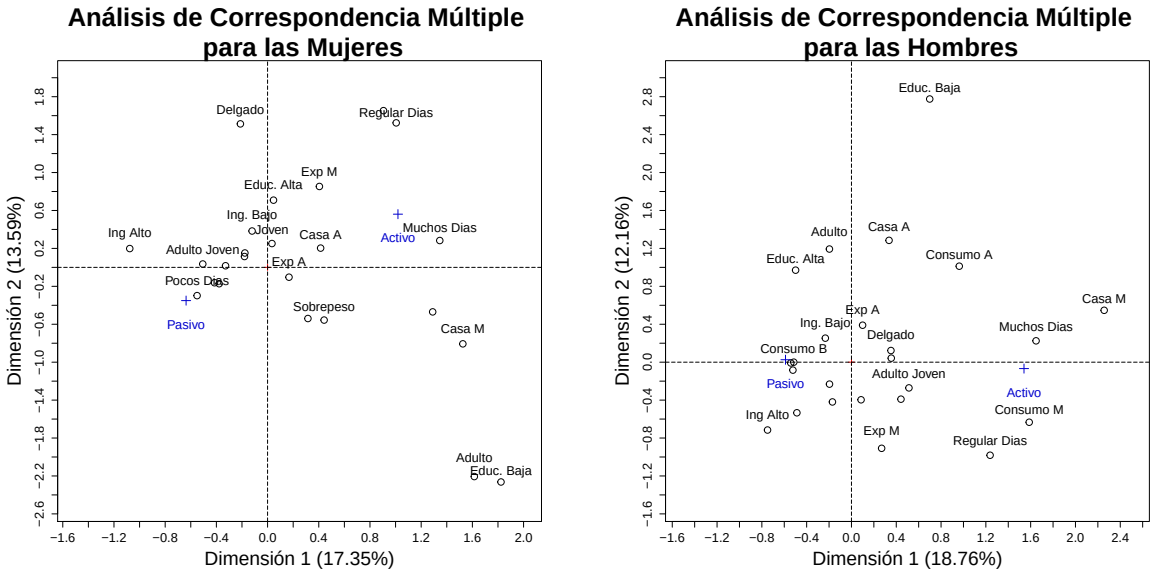

Figura 3.6: Biplots para mujeres y hombres, sin las últimas 4 preguntas.

Cuando se clasifica con respecto a la actividad del fumador, sin importar el género y se realizan los biplots respectivos al análisis de correspondencia se obtiene la Figura 3.7, estos

escenarios se calcularon haciendo muestras separadas con respecto al tipo de fumador, el tamaño de las muestras en cada biplot es diferente, ya que el número de fumadores activos es la mitad del número de fumadores activos. En el escenario de los fumadores activos contra los fumadores pasivos ocurre algo inesperado pero intuitivamente lógico, es decir, véase como la mayoría de las preguntas en cada nivel cambian de posición. Para el caso de los fumadores activos, las mujeres poseen en su mayoría un IMC normal, exposición fuera de casa media, en casa una exposición baja y educación alta, a diferencia de los hombres quienes tienen un mayor consumo de cigarrillos, con problemas de peso (ya sea de bajo peso o de sobrepeso) y con edad promedio con respecto a la muestra.

Para el caso de los fumadores pasivos las características se concentran más en el centro, las cuales son compartidas entre hombres y mujeres. Cabe resaltar que para esta muestra los niveles de cada pregunta en su mayoría parecieran estar en una misma agrupación, es decir la muestra es muy diversa cuando se trata de fumadores pasivos y no hay subgrupos tan marcados como en los casos anteriores. Todos estos comentarios sólo caracterizan a la muestra en cuestión, mas no a la población bajo estudio, y no pretenden establecer factores causales entre ellos.

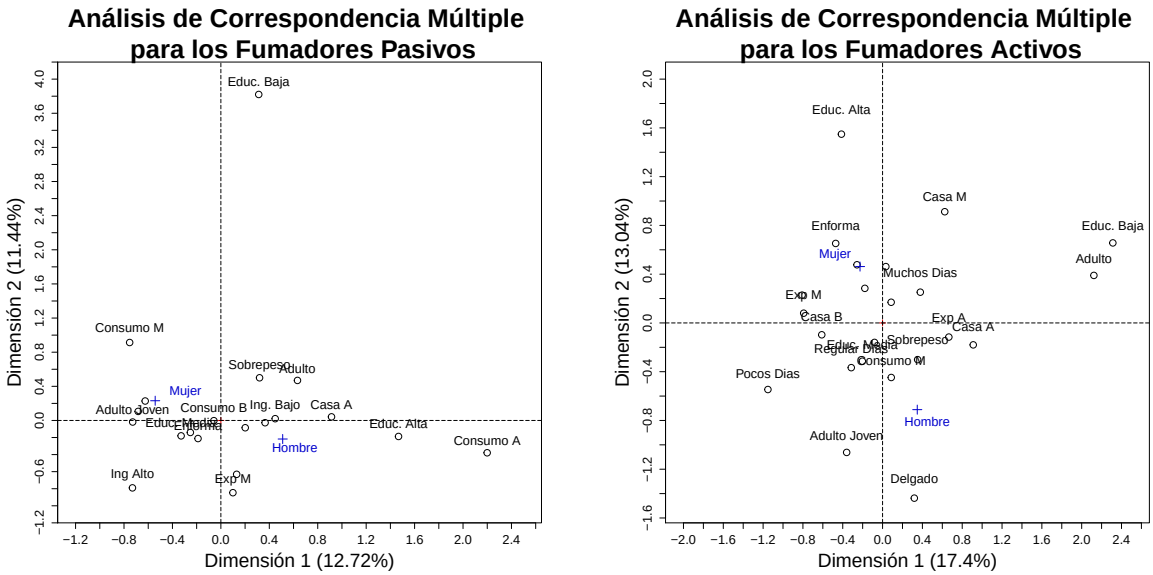

Figura 3.7: Biplots para fumadores pasivos y activos, sin las últimas 4 preguntas.

En el capítulo final se aplicará la teoría con respecto a las técnicas y la metodología estadística de los capítulos anteriores, con la finalidad de obtener algunas conclusiones sólo en el el sentido de lo que la muestra piloto arroja a manera de ejemplificar la metodología y por ende, no pueden considerarse válidas para la población bajo estudio.

Capítulo 4

Aplicación de NMPLSR

En este capítulo se realiza la aplicación de la metodología que se describió en los primeros capítulos. Inicialmente se describirán las variables respuesta, es decir, las variables objetivo para la regresión en términos del problema. Luego se procederá a transformar las variables y a realizar la regresión, para que después se proceda a realizar las predicciones. De igual forma se calcularán los intervalos de percentiles mediante Bootstrap, para los coeficientes de regresión. Al final se presentará la validación cruzada para el modelo desarrollado, y por último se realizará la clasificación con respecto a ciertos grupos de interés.

4.1. Notación y definiciones

En el capítulo 1 y el capítulo 2, se describieron las formas en que se representarán a las variables predictoras X, así como su significado en el ejemplo planteado. Para el ejemplo que se esta abordando, las variables respuesta son aquellas variables externas al cuestionario, y que probablemente tienen la información necesaria para indicar el nivel de exposición en un sujeto. Estas variables respuesta son resultados de análisis químicos en sangre y orina, por lo tanto a las variables respuesta se denominarán *Biomarcadores*. El objetivo primordial de este trabajo de tesis es conocer si existe una relación entre las concentraciones de los biomarcadores y las respuestas al cuestionario, esto es, con la posibilidad de caracterizar por medio del cuestionario a los sujetos con su respectivo nivel de exposición, y posteriormente predecir los niveles de los biomarcadores para un sujeto fuera de la muestra, pero que provenga de la misma población. En otras palabras se desea construir un modelo estadístico que de acuerdo a una muestra base, modele la exposición para cada sujeto, con la finalidad de predecir el nivel de exposición de un nuevo sujeto, siempre y cuando exista una relación entre las variables predictoras y los biomarcadores. En el ejemplo del capítulo anterior, sólo se tomaron 4 tipos de biomarcadores, cabe resaltar que con la metodología antes planteada se pueden tomar más de 4 biomarcadores ($P_2 > 4$), por lo que se tendría una matriz respuesta para los N sujetos como la siguiente:

Tabla 4.1: Matriz de respuestas teóricas para P_2 Biomarcadores.

$Biomarcador_1$	$Biomarcador_2$	...	$Biomarcador_{P_2}$
$y_{1,1}$	$y_{1,2}$	$\cdots$	y_{1,P_2}
$y_{2,1}$	$y_{2,2}$	$\cdots$	y_{2,P_2}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$y_{N,1}$	$y_{N,2}$	$\cdots$	y_{N,P_2}

4.2. Biomarcadores

Se sabe que el humo del cigarrillo contienen numerosas sustancias, que luego pueden medirse en los líquidos corporales de las personas que se exponen a él, tanto de forma activa como pasiva. Por diseño, en el estudio que motivó el trabajo de esta tesis se mide; la concentración sérica de la *Cotinina*, que es un producto intermedio del metabolismo de la nicotina, y el *Monóxido de carbono (CO)*, que es un producto de la combustión con una constante afinidad por la hemoglobina mayor que la del oxígeno, así como la concentración en orina del *Fenantreno* y de *Nitrosamina*[1], ambos grupos de compuestos presentes en el humo del cigarrillo y relacionados con algunas formas de cáncer. Con estas bases, se piensa que las concentraciones de los biomarcadores serán mayores mientras mayor haya sido la exposición al humo del tabaco.

Las concentraciones o los niveles de los biomarcadores anteriores se pueden medir a través de un análisis clínico en sangre o en orina. Los niveles de los biomarcadores se miden de forma continua, dentro de intervalos específicos para cada sustancia. Dado que la metodología que se propone no depende ni de la longitud, ni de los límites de los intervalos, el nivel de cada biomarcador se reescaló al intervalo $[0, 1]$; es decir, si el nivel del biomarcador está entre $[a, b]$, entonces esta información se proyecta o se mapea al intervalo $[0, 1]$. Por lo tanto para los 99 sujetos en la muestra se tiene la matriz de variables respuesta continuas que se presenta en la Tabla 4.2.

4.3. Modelo estadístico semi-métrico

Una vez que se planteó lo anterior, se puede observar que el sujeto i con $i \in \{1, ..., N\}$ tiene P_1 respuestas a preguntas y tiene P_2 niveles de biomarcadores, con $N = 99$, $P_2 = 4$ y $P_1 = 8$, ya que en el análisis de correspondencia se eliminaron las preguntas 9, 10, 11 y 12, por brindar información irrelevante al ejemplo. Por otra parte, la edad y el IMC son medidos en una escala continua, por lo que es mejor que se mantengan en su métrica original; por esta razón se tendrá un modelo semi-métrico. Por consiguiente la matriz de predictores centrados se presenta en la Tabla 4.3.

[1]polycyclic aromatic hydrocarbon nitrosamines

Tabla 4.2: Variables clínicas externas al cuestionario.

ID	Cotinina	CO	Fenantreno	Nitrosamina
1	0.88	0.94	0.92	0.99
2	0.05	0.04	0.09	0.22
3	0.48	0.29	0.63	0.58
4	0.12	0.13	0.30	0.22
5	0.46	0.45	0.59	0.41
6	0.63	0.61	0.66	0.99
7	0.04	0.11	0.24	0.09
8	0.82	0.66	0.66	0.93
9	0.70	0.29	0.31	0.69
10	0.32	0.03	0.36	0.35
11	0.39	0.57	0.31	0.47
⋮	⋮	⋮	⋮	⋮
90	0.96	0.90	0.95	0.86
91	0.88	0.94	0.98	0.66
92	0.37	0.19	0.15	0.07
93	0.42	0.68	0.57	0.47
94	0.92	0.95	0.82	0.75
95	0.89	0.89	0.95	0.76
96	0.28	0.70	0.37	0.68
97	0.51	0.47	0.69	0.66
98	0.36	0.19	0.37	0.16
99	0.53	0.29	0.32	0.45

Mínimos cuadrados parciales no métricos

Para poder implementar PLS, se tiene que proporcionar una métrica a las variables predictoras ordinales, en este caso a las columnas Q_i de la Tabla 4.3. Para realizar dicha tarea, se hará uso de la transformación secundaria de Kruskal (ver apéndice C)[2]. La transformación de Kruskal está implementada en *R* [13], y se recurrirá a una de las aportaciones de la tesis, que es la propuesta del primer componente principal de los biomarcadores (Tabla 4.2) como el vector que dará un orden a las últimas 6 columnas de la Tabla 4.3. Otra aportación que se realiza es la de combinar variables transformadas con variables continuas, es decir, después de transformar las variables ordinales, agregamos los predictores continuos (edad e IMC) y así se obtiene la matriz $\tilde{X}^o$ que es la matriz predictora final para el método PLS, la cual presentamos en la Tabla 4.4.

Cuando se obtiene la matriz $\tilde{X}^o$ se realiza la regresión PLS no métrico (NMPLSR) con toda la muestra, con la finalidad de que se tenga un modelo predictivo para un nuevo individuo

[2]La idea fue tomada de los trabajos del Dr. Giorgio Russolillo.

Tabla 4.3: Matriz de predictores continuos centrados y ordinales.

ID	Edad	IMC	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
1	-0.19	1.51	2	3	2	1	3	2
2	-0.19	-0.13	2	3	1	1	1	1
3	0.34	0.56	2	3	1	2	1	1
4	1.14	-1.16	2	2	1	1	1	1
5	1.41	0.89	2	2	1	1	3	1
6	1.14	-0.33	1	2	3	2	3	2
7	0.07	-0.49	2	2	1	1	3	1
8	2.75	2.20	1	2	3	2	3	3
9	0.34	0.72	2	2	1	1	1	1
10	0.61	-1.31	2	1	1	1	1	1
11	0.34	0.61	2	3	1	2	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	0.07	-0.11	2	3	3	3	3	1
91	-0.19	0.88	2	2	3	3	3	2
92	-0.19	1.02	2	2	1	1	2	1
93	0.07	-0.70	2	3	2	3	2	1
94	-0.46	3.64	2	3	3	1	3	3
95	-1	1.09	2	2	3	3	3	1
96	0.87	-1.37	2	3	2	1	3	1
97	-0.46	-0.46	2	2	2	2	2	1
98	-0.19	-0.59	2	3	3	1	2	1
99	0.61	1.14	2	2	3	2	1	1

que provenga de la misma población que la muestra. El método NMPLSR se implementó con el código del apéndice D.1 con la finalidad de obtener todos los componentes PLS como en el algoritmo de la Figura 2.1. Los componentes PLS obtenidos se muestran en la Tabla 4.5.

El código del apéndice D.1, además de realizar el algoritmo en la Figura 2.1, también calcula la norma de Frobbenius de la matriz $E_{(h)}$, para $h \in \{1, ..., 8\}$. Por ejemplo, para $h = 1$ la norma de Frobbenius de $E_{(1)}$ ($||E_{(1)}||$), que representa la cantidad de información disponible que no pudo explicar el primer componente PLS ($t_{1(1)}$), $||E_{(2)}||$ es la cantidad de información que no pudo explicar el segundo componente PLS ($t_{1(2)}$) y por construcción esta información tampoco fue explicada por $t_{1(1)}$. En otras palabras, si se usan más componentes PLS, menor es la información que no es explicada por éstos. Esto se ilustra en la Figura 4.1.

Dentro del marco de la metodología PLS, una de las cuestiones a tratar es el número adecuado de componentes PLS a considerar en el modelo final. En ocasiones este problema es de suma importancia, ya que se suele buscar el reducir la dimensionalidad del problema. Existen varias formas de calcular el número de componentes PLS, ya que no existe un

Tabla 4.4: Matriz predictora con 2 preguntas continuas centradas y 6 preguntas ordinales transformadas.

ID	Edad	IMC	V_3	V_4	V_5	V_6	V_7	V_8
1	-0.19	1.51	-0.09	0.22	1.07	-0.55	0.64	1.11
2	-0.19	-0.13	-0.09	0.22	-0.61	-0.55	-1.02	-0.51
3	0.34	0.56	-0.09	0.22	-0.61	1.22	-1.02	-0.51
4	1.14	-1.16	-0.09	-0.14	-0.61	-0.55	-1.02	-0.51
5	1.41	0.89	-0.09	-0.14	-0.61	-0.55	0.64	-0.51
6	1.14	-0.33	-0.09	-0.14	1.69	1.22	0.64	1.11
7	0.07	-0.49	-0.09	-0.14	-0.61	-0.55	0.64	-0.51
8	2.75	2.20	-0.09	-0.14	1.69	1.22	0.64	1.25
9	0.34	0.72	-0.09	-0.14	-0.61	-0.55	-1.02	-0.51
10	0.61	-1.31	-0.09	-1.80	-0.61	-0.55	-1.02	-0.51
11	0.34	0.61	-0.09	0.22	-0.61	1.22	-0.44	-0.51
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	0.07	-0.11	-0.09	0.22	1.69	1.81	0.64	-0.51
91	-0.19	0.88	-0.09	-0.14	1.69	1.81	0.64	1.11
92	-0.19	1.02	-0.09	-0.14	-0.61	-0.55	-0.44	-0.51
93	0.07	-0.70	-0.09	0.22	1.07	1.81	-0.44	-0.51
94	-0.46	3.64	-0.09	0.22	1.69	-0.55	0.64	1.25
95	-1.00	1.09	-0.09	-0.14	1.69	1.81	0.64	-0.51
96	0.87	-1.37	-0.09	0.22	1.07	-0.55	0.64	-0.51
97	-0.46	-0.46	-0.09	-0.14	1.07	1.22	-0.44	-0.51
98	-0.19	-0.59	-0.09	0.22	1.69	-0.55	-0.44	-0.51
99	0.61	1.14	-0.09	-0.14	1.69	1.22	-1.02	-0.51

acuerdo común en la literatura de PLS, por ejemplo si se opta por seguir una forma empírica, bastará con tener el número de componentes PLS que expliquen al menos el 80 % de la información disponible. Esto se puede ver en la Figura 4.2, por lo tanto bastaría con que se tomen de 4 a 5 componentes PLS. Sin embargo el numero de columnas en de X no es grande, así que por el momento se tomarán los 8 componentes PLS.

Tabla 4.5: La columna h de la matriz $\mathbf{t_1}$, es el componente h PLS ($t_{1(h)}$).

	$t_{1(1)}$	$t_{1(2)}$	$t_{1(3)}$	$t_{1(4)}$	$t_{1(5)}$	$t_{1(6)}$	$t_{1(7)}$	$t_{1(8)}$
1	1.43	0.70	0.16	-0.33	-1.44	0.36	-0.17	-0.34
2	-1.21	-0.48	-0.19	0.16	0.02	0.20	-0.24	0.01
3	0.03	-0.64	0.10	0.96	0.57	-0.24	-0.33	0.54
4	-1.21	-1.30	0.06	-0.49	0.69	0.03	-0.11	0.12
5	0.09	-0.53	1.34	0.14	-0.29	-0.18	0.14	-0.20
6	2.37	-0.14	-0.17	-0.66	0.42	-0.16	-0.05	-0.04
7	-0.75	0.12	0.42	-0.07	0.21	-0.23	0.15	-0.39
8	3.71	-0.86	1.12	-0.24	-0.74	-0.06	-0.11	0.29
9	-0.81	-0.84	0.09	0.10	-0.66	0.07	-0.16	0.20
10	-1.57	-1.51	-0.91	-1.51	-0.70	-0.69	0.31	0.48
11	0.23	-0.34	0.36	1.02	0.61	-0.33	-0.23	0.37
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	1.86	-0.03	-0.60	0.64	0.84	-0.19	0.14	-0.54
91	2.68	0.61	-0.57	0.05	-0.38	-0.30	-0.12	0.11
92	-0.67	-0.26	0.16	0.29	-0.91	-0.02	-0.08	0.01
93	0.96	-0.50	-0.89	0.54	1.22	-0.16	-0.09	-0.09
94	2.46	0.83	0.20	0.07	-2.98	0.54	-0.20	-0.35
95	1.93	0.44	-0.99	0.86	-0.46	-0.33	0.19	-0.44
96	0.17	-0.50	-0.05	-0.56	0.84	0.27	0.23	-0.98
97	0.57	-0.36	-1.20	0.22	0.36	-0.17	0.02	-0.16
98	0.13	-0.60	-1.15	-0.45	-0.19	0.58	0.06	-0.87
99	1.56	-1.28	-0.92	0.33	-0.55	0.12	-0.05	0.01

4.3.1. Regresión simple

Cuando se obtiene la matriz $\mathbf{t_1}$, se realiza una regresión usual por mínimos cuadrados para cada biomarcador en $\mathbf{Y} = (Y_{Cot}, Y_{CO}, Y_F, Y_N)$. Con base en la ecuación (1.6) se obtienen las siguientes estimaciones de los coeficientes, las cuales presentamos en la Tabla 4.6.

Tabla 4.6: Coeficientes de regresión de mínimos cuadrados de los biomarcadores $\mathbf{Y}$ sobre $\mathbf{t_1}$.

β_{MC}	$t_{1(1)}$	$t_{1(2)}$	$t_{1(3)}$	$t_{1(4)}$	$t_{1(5)}$	$t_{1(6)}$	$t_{1(7)}$	$t_{1(8)}$
β_{Cot}	0.53	0.25	0.04	0.19	0.06	0.13	-0.05	-0.05
β_{CO}	0.48	0.33	0.12	0.24	0.08	0.23	0.15	0.05
β_{Fenan}	0.52	0.25	0.09	0.02	0.12	0.00	-0.12	0.08
β_{Nitro}	0.54	0.18	0.18	0.09	0.07	0.27	0.31	0.18

En la sección 1.3 se resalta la importancia de recuperar el modelo original $Y = X\beta^*$, por lo tanto se parte del modelo $Y = \mathbf{t_1}\beta$ y con el código del apéndice D.1, se puede obtener el

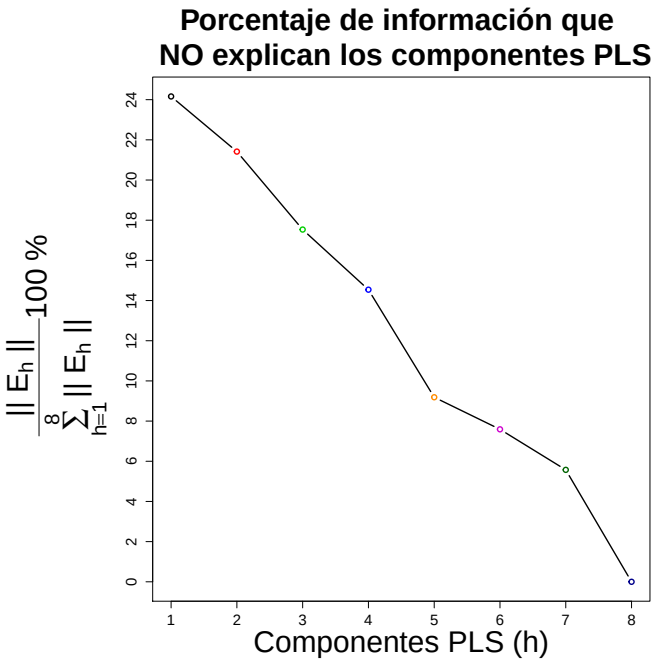

Figura 4.1: A mayor cantidad de componentes PLS calculados, menor es la información que no se puede explicar.

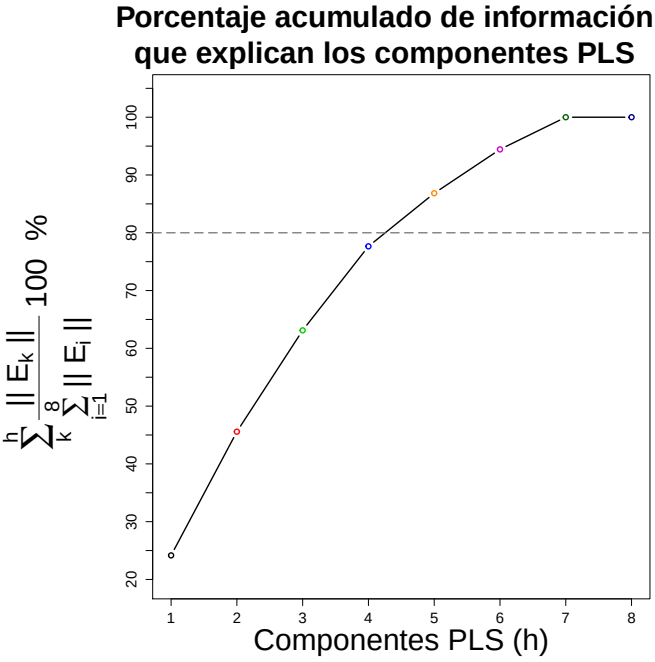

Figura 4.2: Ilustración de la regla empírica del 80 %.

modelo $Y = X\beta^*$ para cada biomarcador, ya que el código está basado en la ecuación (1.8). Si se consideran los 8 componentes PLS, entonces se obtienen los resultados de la Tabla 4.7.

En base a la Tabla 4.7 se recuperan los valores reales de los biomarcadores, es decir, hacer predicción dentro de la muestra. Éstos resultados se presentan en la Tabla 4.8.

Tabla 4.7: Coeficientes de regresión de los indicadores de exposición **Y** sobre X, con todos los componentes PLS.

Indicador	Edad	IMC	V_3	V_4	V_5	V_6	V_7	V_8
Cotinina	0.031	0.207	0.141	0.457	0.191	0.332	0.355	0.197
CO	0.026	0.205	0.636	0.546	0.088	0.331	0.425	0.179
Fenantreno	0.093	0.137	0.101	0.329	0.106	0.363	0.332	0.329
Nitrosamina	0.169	0.212	0.962	0.340	0.151	0.288	0.343	0.224

Tabla 4.8: Predicción dentro de la muestra con 8 componentes PLS.

ID	Cotinina	CO	Fenantreno	Nitrosamina
1	0.86	0.77	0.75	0.75
2	-0.70	-0.72	-0.74	-0.78
3	0.04	0.02	0.04	-0.04
4	-1.04	-1.11	-0.88	-0.90
5	-0.02	0.04	-0.02	0.15
6	1.06	0.85	1.21	1.06
7	-0.35	-0.29	-0.34	-0.37
8	1.66	1.46	1.75	1.90
9	-0.68	-0.73	-0.69	-0.63
10	-1.85	-2.06	-1.50	-1.59
11	0.26	0.28	0.24	0.17
⋮	⋮	⋮	⋮	⋮
90	1.12	0.97	0.94	0.86
91	1.47	1.27	1.46	1.27
92	-0.42	-0.43	-0.51	-0.46
93	0.49	0.33	0.43	0.27
94	1.44	1.30	1.12	1.28
95	1.17	1.00	0.88	0.81
96	-0.02	-0.12	-0.08	-0.05
97	0.17	-0.02	0.08	-0.06
98	-0.15	-0.38	-0.37	-0.34
99	0.44	0.16	0.27	0.35

Ahora se realiza el procedimiento anterior, teniendo en cuenta diferentes números de componentes PLS. A continuación se muestra el ajuste del modelo estadístico en la Figura 4.3. En la implementación del modelo estadístico con los 99 sujetos, se puede recuperar los

valores de los biomarcadores centrados dentro de la muestra, una vez hecho esto véase como la predicción parece ser razonable, en la Figura 4.3.

Por otra parte, en cuanto al problema del número de componentes PLS adecuado, una prueba gráfica que brinda una idea del número de componentes PLS sensatos para el modelo es la siguiente: ya que se obtuvienen las estimaciones para los 99 sujetos en la muestra, se grafica la norma de Frobbenius de la matriz de diferencias entre los indicadores predichos y los indicadores reales. El resultado de este método se ilustra en la Figura 4.4.

Sa analiza la Figura 4.4, y se nota que tomar sólo el primer componente implica mayor error en la predicción, a diferencia de que se tome 7 u 8 componentes. Esta medición del error de predicción es subjetiva, ya que las predicciones son de los datos de aprendizaje del modelo. Esto se puede contrastar con el criterio empírico en el cual bastaba tomar 4 ó 5 componentes PLS. Para el ejemplo, se pueden tomar libremente los componentes necesarios, ya que en este caso la dimensión de las variables predictoras no es grande.

4.4. Intervalos de confianza Bootstrap

Cada vez que se realiza alguna predicción o estimación, es útil calcular intervalos de confianza. Los intervalos de confianza brindan mayor información en comparación con una estimación puntual, puesto que miden la incertidumbre de la estimación y proporcionan un intervalo que captura al parámetro real con cierta probabilidad. En este caso las estimaciones que se hicieron fueron hechas con mínimos cuadrados (de forma numérica), de tal forma que no se esta haciendo ningún supuesto distribucional. Como se mencionó en el capítulo 3, los intervalos de confianza Bootstrap son razonables para esta situación. La matriz a la que se le aplicó la metodología PLS se presenta en la Tabla 4.4. La matriz de predictores centrada transformada (Tabla 4.4) es la matriz con la que se hará el muestreo con reemplazo.

De la misma forma que se usaron los predictores centrados, también se usarán las variables respuesta centradas, es decir, los biomarcadores centrados y codificados (Tabla 4.2). Para calcular los intervalos de confianza por percentiles, se hace uso de la función `boot` en *R* [13], el código se encuentra en el apéndice D.3. Para los intervalos de confianza se realizaron $\mathbf{R} = 10,000$ réplicas bootstrap, para obtener diferentes valores del estadístico de interés para cada biomarcador, los histogramas y la gráfica de cuantiles se pueden observar en la Figura 4.5. Con las réplicas bootstrap se obtiene el estadístico de interés β_i^*, el cual es el coeficiente de regresión de la pregunta i para cierto biomarcador. Estos resultados se muestran en la Tabla 4.9 y la Tabla 4.10.

Tabla 4.9: Tabla de intervalos de confianza para Cotinina y Monóxido de carbono.

Cotinina		
i	β^*_{il}	β^*_{iu}
Edad	-0.09	0.18
IMC	0.06	0.37
V_3	-4.91	6.37
V_4	-0.13	2.38
V_5	-0.55	0.50
V_6	-0.13	0.83
V_7	0.13	1.28
V_8	-0.30	0.68

CO		
i	β^*_{il}	β^*_{iu}
Edad	-0.11	0.19
IMC	0.05	0.37
V_3	-1.75	3.56
V_4	0.35	2.66
V_5	-1.11	0.40
V_6	-0.10	1.11
V_7	0.27	1.41
V_8	-0.61	0.65

Tabla 4.10: Tabla de intervalos de confianza para Fenantreno y Nitrosamina.

Fenantreno		
i	β^*_{il}	β^*_{iu}
Edad	-0.07	0.26
IMC	-0.01	0.29
V_3	-5.36	7.21
V_4	-0.25	2.33
V_5	-0.77	0.38
V_6	0.01	0.97
V_7	0.14	1.09
V_8	-0.03	0.71

Nitrosamina		
i	β^*_{il}	β^*_{iu}
Edad	0.04	0.30
IMC	0.06	0.37
V_3	-0.10	3.59
V_4	0.02	2.59
V_5	-0.60	0.47
V_6	-0.15	0.84
V_7	0.05	1.01
V_8	-0.19	0.71

De los resultados que se obtuvieron, nótese que la variable predictora que corresponde al nivel educativo más alto alcanzado (V_3) no es significativa y tiene los intervalos de confianza Bootstrap más amplios que cualquier otra variable predictora. Por el contrario la variable que corresponde a la frecuencia con la que alguien fuma en su presencia en lugares cerrados (V_7) es significativa para los 4 biomarcadores con una longitud promedio de una unidad. En términos de significancia estadística o de mejor impacto al predecir los biomarcadores se encuentra el IMC (excepto únicamente en el Fenantreno). Este resultado intuitivamente puede llevar a pensar que el IMC y la exposición en lugares cerrados son factores importantes para los biomarcadores. Algo peculiar que ocurre en las tablas anteriores es el alto sesgo que tiene la pregunta nivel de ingresos familiar mensual en todos los biomarcadores, y eso en parte le ayuda a conseguir que sea significativa en la Nitrosamina y el CO una posible causa de este acontecimiento es que posiblemente la etiqueta que se le puso no fue la mejor para el algoritmo.

El hacer una regresión PLS es similar a realizar una regresión usual, ya que se construye

un modelo lineal y el cálculo de los coeficientes es el mismo. De esa misma forma, se puede ver como los coeficientes pudieran brindarnos información sobre la relación que existe entre las variables predictoras y las variables respuesta. Nótese como los coeficientes impactan más en unos biomarcadores, y menos en otros, por otro lado la literatura de PLS recomienda el uso de la metodología Bootstrap para el cálculo de intervalos de confianza, ya que no se cuenta con teoría asintótica suficiente para el cálculo de dichos intervalos. El hecho de que los intervalos de confianza Bootstrap contengan al valor cero, a primera vista indica que no aportan información al modelo, pero no se tiene que hacer a un lado que el modelo tiene gran variabilidad en los coeficientes y que el objetivo del modelo es el clasificar a los sujetos en la muestra.

4.5. Validación cruzada

Cuando se realiza un análisis estadístico existen dos tipos de datos: los datos de entrenamiento o de aprendizaje y los datos de prueba, dicha clasificación de datos se usa cuando se quiere implementar la *validación cruzada* o *cross-validation*. La validación cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la muestra tomada y de la partición entre datos de entrenamiento y los datos de prueba. En otras palabras, la validación cruzada evalúa el desempeño de un modelo estadístico. Desde otra perspectiva la validación cruzada es una manera de medir el ajuste de un modelo a un conjunto hipotético de datos de prueba, cuando no se puede disponer del conjunto explícito de datos de prueba.

4.5.1. Validación cruzada dejando uno fuera

La *validación cruzada dejando uno fuera* o *leave-one-out cross-validation* implica que se separan los datos de forma que para cada iteración se tenga un solo individuo para los datos de prueba y el resto de la muestra conforma los datos de entrenamiento. La evaluación del método está en términos de algún criterio, y lo que se tome como medida clasificadora dependerá del contexto del problema. En este tipo de validación cruzada el error es muy bajo, en cambio, a nivel computacional es muy costoso, puesto que se tiene que realizar un elevado número de iteraciones, tantas como N individuos que se tengan en la muestra y para cada uno analizar los datos tanto de entrenamiento como de prueba.

Para visualizar el ajuste en el método de validación cruzada dejando uno afuera, véase la Figura 4.6. Nótese que el ajuste en la mayoría de las iteraciones de la validación cruzada es bastante sensato, e inclusive son menos estables a los resultados obtenidos cuando se usa toda la muestra para obtener el modelo estadístico (ver Figura 4.3).

Análogamente al modelo con la muestra completa, cuando se realiza la validación cruzada también se puede medir el desempeño de la predicción con respecto al número de componentes PLS tomados. Este desempeño como los anteriores, se mide con la norma de Frobbenius

Tabla 4.11: Predicción de la validación cruzada dejando uno fuera.

ID	Cotinina	CO	Fenantreno	Nitrosamina
1	0.74	0.63	0.67	0.59
2	0.28	0.19	0.22	0.19
3	0.50	0.46	0.43	0.37
4	0.15	0.04	0.15	0.14
5	0.53	0.51	0.48	0.51
6	0.78	0.62	0.81	0.64
7	0.46	0.42	0.40	0.38
8	0.94	0.81	0.99	0.87
9	0.24	0.17	0.23	0.22
10	-0.41	-0.44	-0.27	-0.26
11	0.61	0.53	0.58	0.47
⋮	⋮	⋮	⋮	⋮
90	0.80	0.68	0.65	0.60
91	0.88	0.72	0.86	0.73
92	0.35	0.29	0.30	0.30
93	0.61	0.43	0.50	0.43
94	0.89	0.72	0.76	0.79
95	0.78	0.64	0.59	0.57
96	0.54	0.34	0.41	0.35
97	0.46	0.31	0.37	0.30
98	0.40	0.26	0.23	0.33
99	0.49	0.32	0.41	0.39

de la diferencia entre los indicadores predichos y los reales, esto se puede ver en la Figura 4.7, que a pesar que se ve creciente la curva, dicho crecimiento es relativamente bajo ya que el valor oscila entre 14 y 15 de forma suave, por lo que es indiferente el número de componentes PLS a considerar.

Aunque los ajustes son similares en ambos escenarios del método, el desempeño del número de componentes PLS a considerar cambia, es decir, el considerar 1 o más componentes visualmente no hace la diferencia. Por otra parte cuando se predice a un sujeto fuera de muestra tenemos más variación en las estimaciones, a diferencia de predecir dentro de muestra. Sin embargo por el contexto del ejemplo en ambos casos buscamos que no subestime las predicciones. Más adelante se hablará sobre otra forma de medir y comparar el desempeño de usar diferentes número de componentes PLS.

4.6. Clasificación

Uno de los usos principales para la metodología que se presenta, es el de clasificación de los sujetos de acuerdo a su nivel de exposición a cada biomarcador. Esto es gracias a que el modelo predice las respuestas de un sujeto dada una muestra de entrenamiento, con esta predicción y ciertos criterios, se entrena al algoritmo para clasificar a partir de dichos criterios o reglas duras. Una vez que se establecieron límites o márgenes para los niveles de Cotinina, CO, Fenantreno y Nitrosamina, se puede medir la intensidad del riesgo que tienen los sujetos a contraer enfermedades pulmonares, y se puede clasificar a la población en los niveles 1, 2 y 3 (*bajo, medio y alto*), los cuales se muestran en la Tabla 4.12.

Tabla 4.12: Matriz de clasificación real.

Indicador	Bajo	Medio	Alto
Cotinina	27	44	28
CO	31	43	25
Fenantreno	29	45	25
Nitrosamina	28	44	27

Por otra parte si se hace la clasificación con los mismos puntos de corte basándose en las predicciones, se puede tener una idea de que tan bueno o malo es la clasificación que se hace. Para este caso se realizó la clasificación para todos los componentes, donde los escenarios más sobresalientes ocurrieron cuando usan 1 ó 2 componentes, entonces se obtienen la Tabla 4.13 y la Tabla 4.14.

Tabla 4.13: Matriz de clasificaciones con 1 componente PLS

Biomarcador	Bajo	Medio	Alto
Cotinina	7	75	17
CO	11	76	12
Fenantreno	13	74	12
Nitrosamina	7	78	14

Tabla 4.14: Matriz de clasificaciones con 2 componentes PLS.

Biomarcador	Bajo	Medio	Alto
Cotinina	16	58	25
CO	18	59	22
Fenantreno	21	56	22
Nitrosamina	11	68	20

Con las tablas de clasificación se puede pensar en medir la efectividad de las predicciones. Para medir la efectividad de las clasificaciones se hace un cruce entre las tablas con respecto

observaciones donde fue clasificada, dado que tiene una clasificación real con respecto a un biomarcador determinado. De esta forma se obtiene la tabla cruzada de 3×3, donde los renglones son la clasificación real y las columnas pertenecen a la clasificación predicha, tal y como se ilustra en la Tabla 4.15.

Tabla 4.15: Tabla de clasificaciones teórica, dado un biomarcador.

	Biomarcador		
	1	2	3
1	d_1	FP_1	FP_2
2	FN_1	d_2	FP_3
3	FN_2	FN_3	d_3

Por ejemplo el sujeto i está en la clasificación 2 de acuerdo al biomarcador j (renglón 2 de la tabla 3×3), al realizar la clasificación, si este sujeto cae en la clasificación correcta (columna 2), el modelo está prediciendo bien, ahora si el sujeto cae en el grupo 3 (columna 3), el modelo está sobreestimando la clasificación y el sujeto tendría que ir a hacerse análisis para verificar su clasificación. Por el contrario en el peor escenario, si el sujeto cae en la categoría 1 (columna 1) podría no hacerse una segunda revisión y correr un riesgo sin saberlo. Estos diferentes escenarios brindan una idea para medir la precisión del modelo de acuerdo a cierto número de componentes PLS considerados y respecto a cierto biomarcador. Por lo tanto partiendo del ejemplo anterior los elementos en la diagonal (d_i) son clasificaciones correctas, los elementos arriba de la diagonal son falsos positivos (FP_i) y por último los elementos restantes son falsos negativos (FN_i). Habiendo definido los elementos anteriores, lo que deseamos es maximizar los elementos d_i, pero minimizando el número de falsos negativos, lo que equivale a también maximizar la suma de la diagonal y en segundo plano maximizar la suma de falsos positivos. Sea $D = \sum_i d_i$ y $FP = \sum_i FP_i$, si se grafica cada valor de estas sumas en una escala del 0 al 1, con respecto a las predicciones que se hicieron con diferentes números de componentes PLS, entonces se obtienen las gráficas en la Figura 4.8 y la Figura 4.9.

Si se desea maximizar la suma de clasificaciones correctas basta con elegir el punto más alto de dicha curva, pero un segundo criterio para medir la efectividad de los componentes es el de maximizar la suma de los falsos positivos. De las gráficas en la Tabla 4.8 y la Tabla 4.9, y con base a lo que se describió anteriormente, las predicciones con 2 componentes son bastantes razonables y coherentes.

4.6.1. Clasificaciones por categorías

Como se mencionó anteriormente, además de tener variables predictoras y variables predichas, en el ejemplo se tienen otras características de los sujetos, estas son el género o el sexo al nacer y si son fumadores pasivos o activos. Esta información paralela al modelo ayudará a crear más escenarios, con el objetivo de tener otra perspectiva de los datos, sus resultados y poder contrastar los grupos dentro de cada escenario.

Fumadores pasivos vs fumadores activos

La siguiente clasificación tiene el objetivo de poder comparar la efectividad del modelo entre fumadores activos y pasivos. A continuación se muestran las clasificaciones que se obtuvieron tanto para los fumadores pasivos en la Tabla 4.16, como para los fumadores activos en la Tabla 4.17.

Tabla 4.16: Tabla de clasificaciones con respecto a los fumadores **pasivos**

Cotinina

	1	2	3
1	9	16	0
2	5	24	2
3	1	9	0

CO

	1	2	3
1	10	16	0
2	7	25	0
3	0	8	0

Fenantreno

	1	2	3
1	11	14	0
2	7	25	0
3	2	7	0

Nitrosamina

	1	2	3
1	6	18	0
2	4	29	2
3	0	9	0

Tabla 4.17: Tabla de clasificaciones con respecto a los fumadores **activos**

Cotinina

	1	2	3
1	1	1	0
2	0	6	7
3	0	2	16

CO

	1	2	3
1	1	3	1
2	0	4	7
3	0	3	14

Fenantreno

	1	2	3
1	1	2	1
2	0	6	7
3	0	2	14

Nitrosamina

	1	2	3
1	1	2	1
2	0	5	6
3	0	5	13

Ya que el número de pasivos es el doble que el número de fumadores activos en la muestra, los puntajes en la Tabla 4.16 tienen que ser divididos entre dos para poder ser comparados con los puntajes de la Tabla 4.17. Una vez realizado lo anterior, se comparan ambas tablas entonces, se dice que el modelo clasifica mejor a los fumadores activos, lo cual pudiera deberse a que su exposición al tabaquismo es constante (los hábitos de tabaquismo son más marcados) y en cuanto a los fumadores pasivos probablemente su nivel de riesgo está relacionado con otros factores no considerados en el modelo o el cuestionario.

Análisis conjunto de mujeres vs hombres

Análogamente al escenario de fumadores, ahora se realizará el mismo procedimiento con la variable sexo al nacer. El modelo predictivo se realizó con toda la muestra, y por tal motivo estos contrastes son para comparar el rendimiento de las predicciones con respecto al sexo al nacer de cada sujeto. A continuación se muestran las clasificaciones tanto para las mujeres en la Tabla 4.18, como para los hombres en la Tabla 4.19.

Tabla 4.18: Tabla de clasificaciones con respecto a las **mujeres**

Cotinina

	1	2	3
1	8	9	0
2	2	17	6
3	0	3	7

CO

	1	2	3
1	9	10	0
2	2	16	3
3	0	4	8

Fenantreno

	1	2	3
1	8	7	0
2	4	17	2
3	1	4	9

Nitrosamina

	1	2	3
1	6	12	1
2	1	18	3
3	0	4	7

Tabla 4.19: Tabla de clasificaciones con respecto a los **hombres**

Cotinina

	1	2	3
1	2	8	0
2	3	13	3
3	1	8	9

CO

	1	2	3
1	2	9	1
2	5	13	4
3	0	7	6

Fenantreno

	1	2	3
1	4	9	1
2	3	14	5
3	1	5	5

Nitrosamina

	1	2	3
1	1	8	0
2	3	16	3
3	0	10	6

Este análisis tiene por objetivo evaluar las predicciones de acuerdo al sexo, esto se hizo de forma conjunta, es decir, todos los sujetos entran en el mismo entrenamiento y una vez predichos sus valores se clasifican de acuerdo al sexo y para luego medir la efectividad en cada grupo. Al comparar los resultados se tiene que el grupo mejor clasificado, pertenece al grupo de las mujeres. Este resultado claramente marca énfasis que existe una diferencia entre las reacciones de los hombres y mujeres con respecto al tabaquismo.

Así como los escenarios anteriores, si se tiene suficiente información se puede hallar una clasificación adicional al grupo de estudio entonces, se realizan más escenarios y se comparar los contrastes dentro de los escenarios.

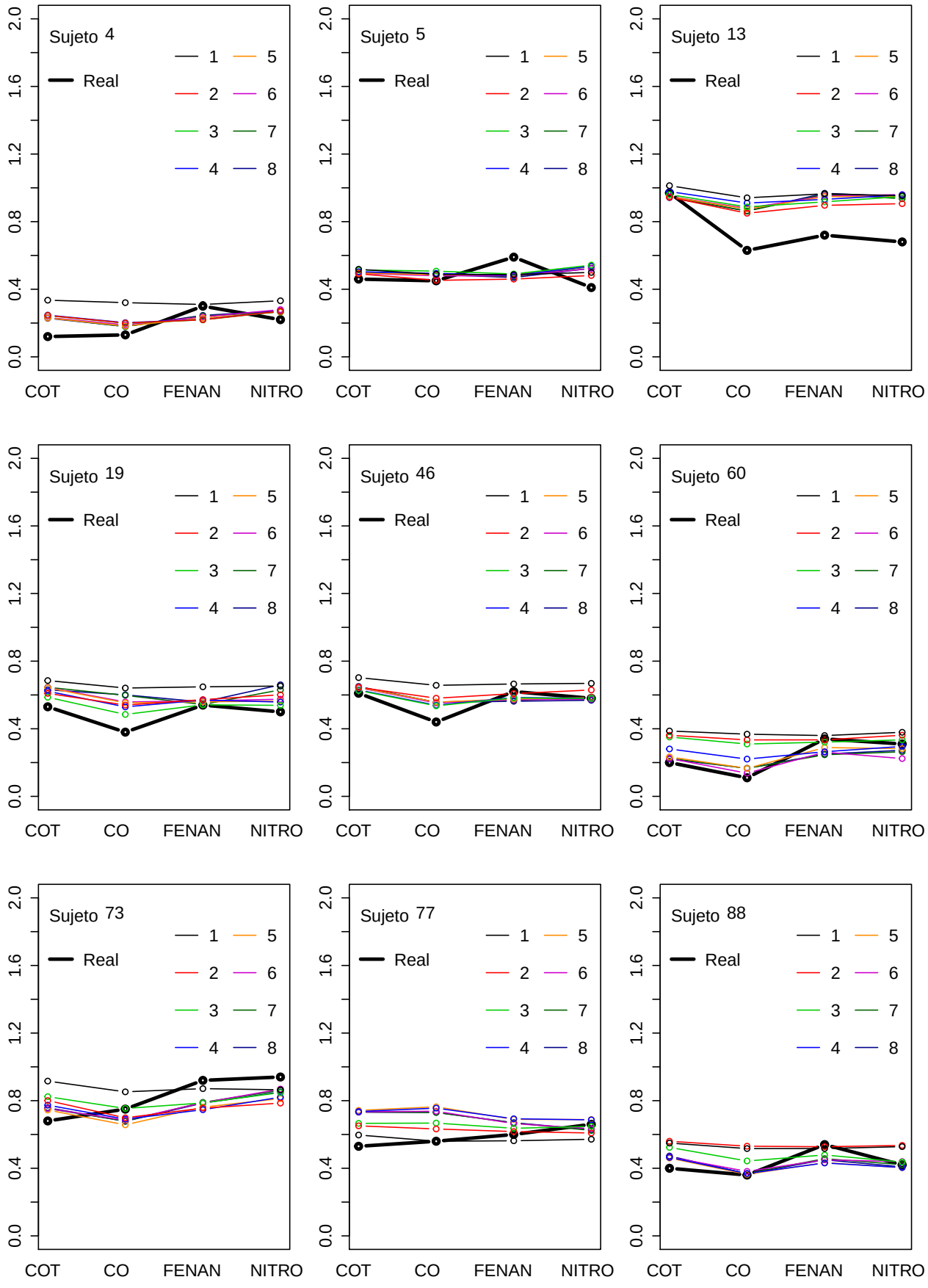

Figura 4.3: Ajuste de los Biomarcadores codificados estandarizados de algunos sujetos, tomando en cuenta diferente número de componentes PLS.

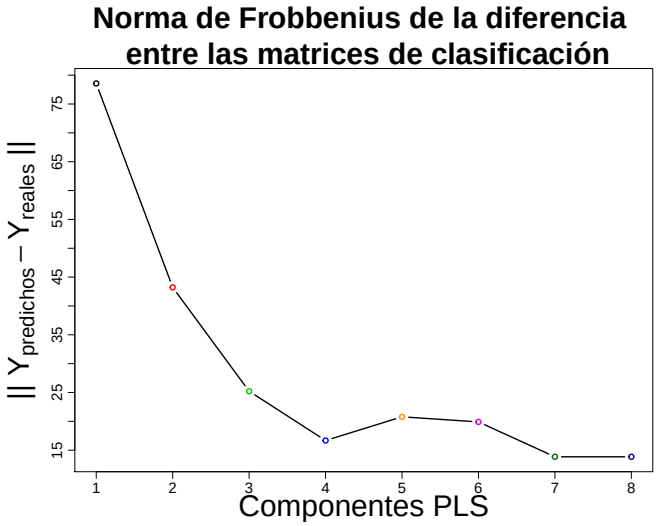

Figura 4.4: Norma de Frobbenius de la diferencia entre los indicadores predichos y los indicadores reales.

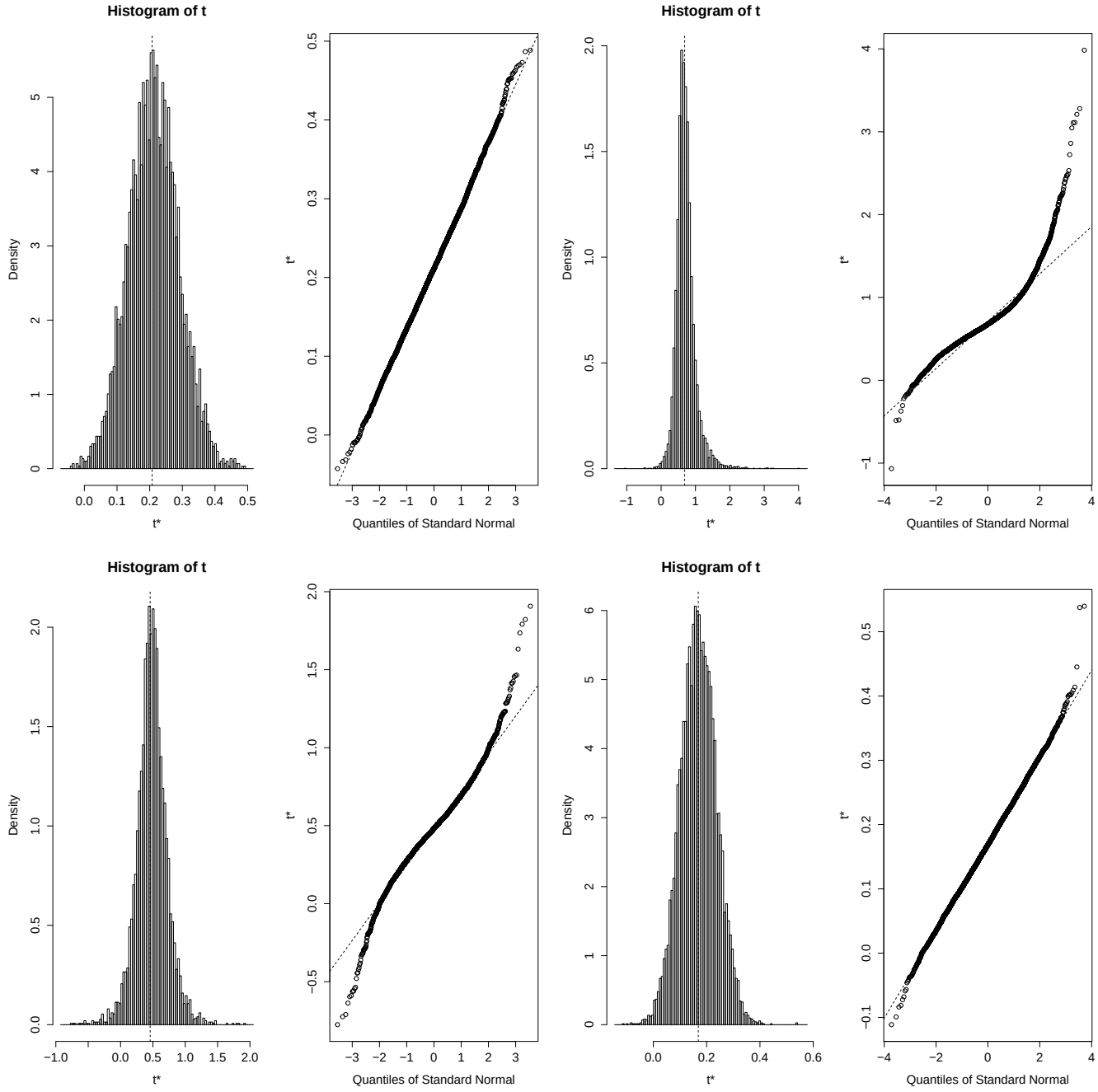

Figura 4.5: Histogramas y gráficas de quantiles normales de la simulación Bootstrap para cada biomarcador. En la esquina superior izquierda está el biomarcador 1 de la variable IMC, a la derecha de este se encuentra el biomarcador 2 con la variable V_7. Abajo a la izquierda el biomarcador 3 con la variable V_6 y al final el biomarcador 4 con la variable Edad.

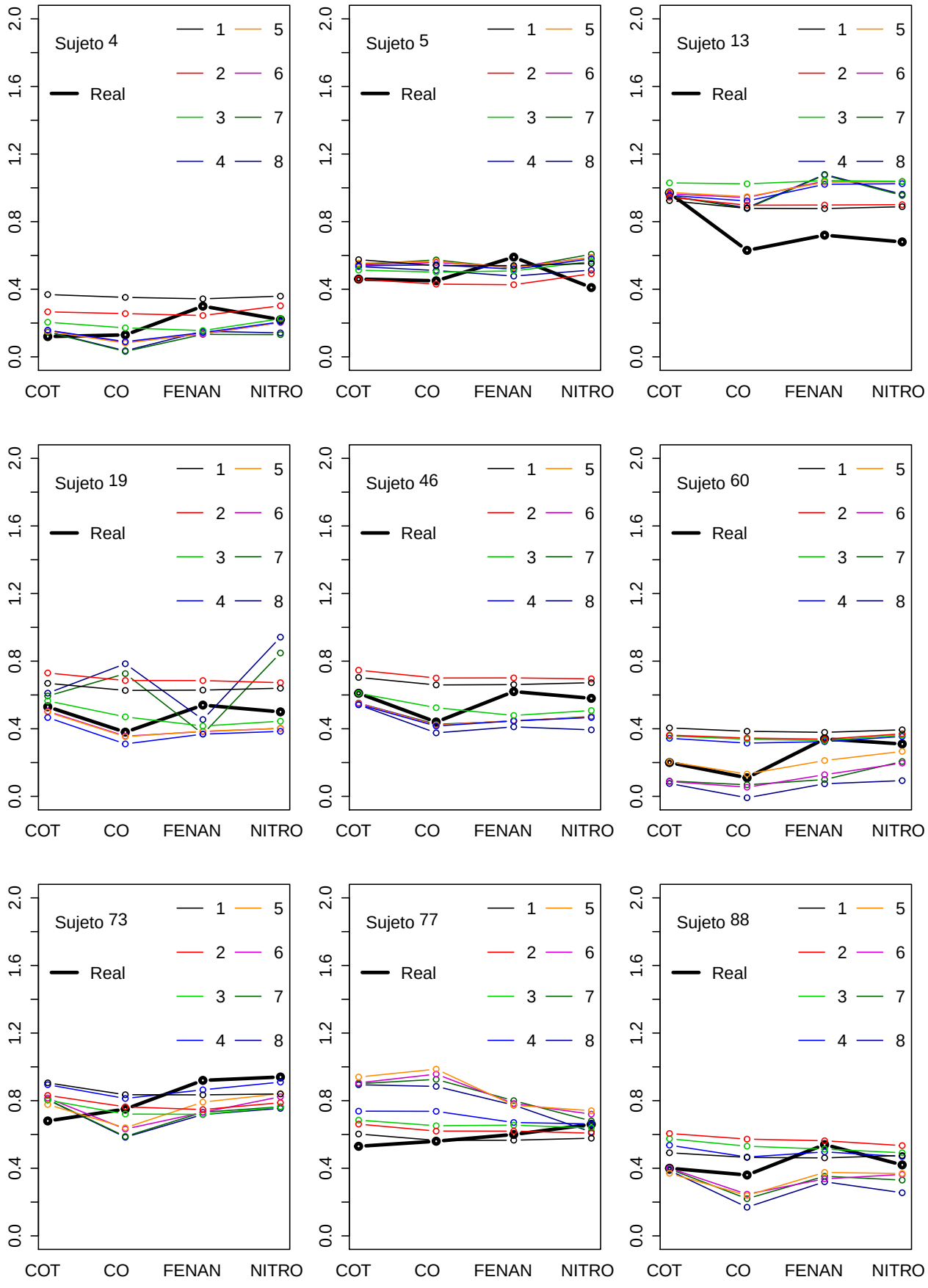

Figura 4.6: Validación cruzada dejando uno fuera tomando en cuenta diferente número de componentes PLS.

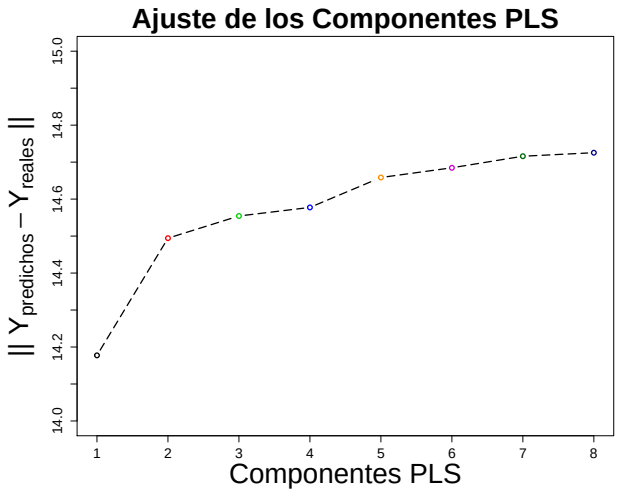

Figura 4.7: Validación cruzada dejando uno fuera tomando en cuenta diferente número de componentes PLS.

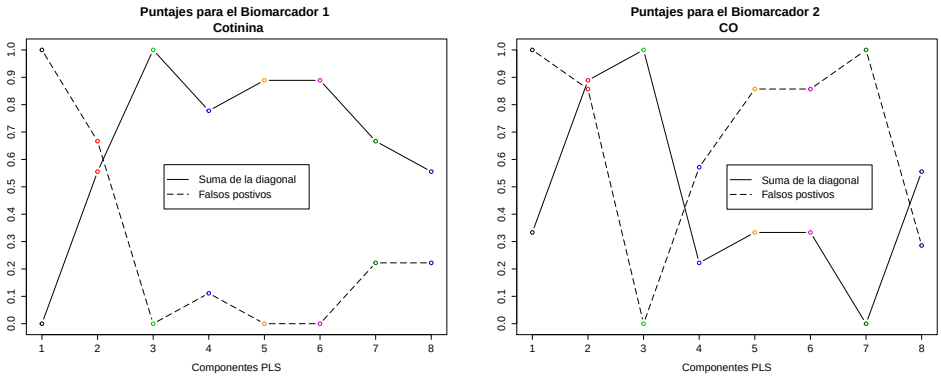

Figura 4.8: Validación cruzada dejando uno fuera tomando en cuenta diferente número de componentes PLS, para el biomarcador 1 y 2.

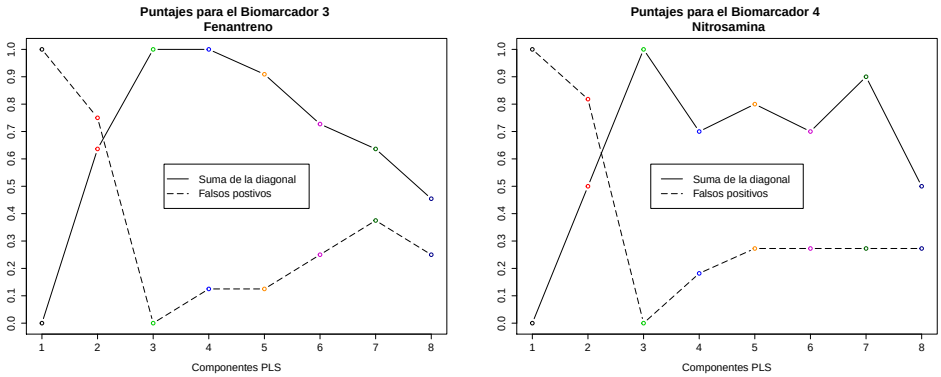

Figura 4.9: Validación cruzada dejando uno fuera tomando en cuenta diferente número de componentes PLS, para el biomarcador 3 y 4.

Capítulo 5

Conclusiones

La motivación del trabajo de tesis es la de explotar y obtener la mejor calidad de información disponible con respecto al proyecto del Análisis Toxicológico y Toxicogenómico para un grupo expuesto al tabaco. Como lo se mencionó al principio, la tesis está basada en una muestra piloto de la cual se hizo un análisis exploratorio con la finalidad de conocer con mayor detalle y profundidad la información disponible. La muestra piloto es una muestra diversa en cuanto al IMC y balanceada en cuanto al sexo al nacer se refiere, sin embargo con respecto a la proporción de fumadores activos contra los fumadores pasivos ocurre lo contrario, la muestra en este sentido está desbalanceada y esto puede ocasionar algunos problemas en cuanto a comparar estos subconjuntos. El análisis de correspondencia brinda un panorama de la importancia estadística de cada nivel en las variable predictoras. Con respecto al análisis no métrico de regresión PLS se tiene que la variable que mide la frecuencia con la que un tercero fuma en lugares cerrados es importante para los cuatro biomarcadores, seguida del IMC el cual únicamente no fue significativo para el fenantreno. Esta significancia de la que se habla se reduce al alto impacto que tienen estas variables para predecir los niveles de los biomarcadores. Sin embargo con todos los resultados anteriores, cabe destacar que como en cualquier campo de aplicación siempre tiene que haber un guía, "expertise" o en este caso un experto en el área de salud, con el objetivo de no perder información valiosa o variables importantes en el proceso de selección de variables para explicar e interpretar los resultados de los análisis.

El trabajo en general se puede resumir en los siguientes pasos:

1. Se tienen 2 conjuntos de variables uno en escala continua y el otro en escala ordinal, de forma conjunta integran las variables predictoras X. Por otro lado se tienen variables respuesta Y en escala continua.

2. Se estandarizan las variables respuesta Y para realizar el análisis de componentes principales, en este caso se hizo vía NIPALS y el primer componente resultante define el orden del escalamiento en la transformación de Kruskal.

3. Se aplica la transformación de Kruskal a la parte ordinal de las variables predictoras con base en el primer componente principal (obtenido anteriormente). Con el resultado de la transformación de Kruskal, junto con las predictoras continuas estandarizadas obtenemos X^o, el cual se usa para la regresión PLS.

4. Con la matriz X^o y las variables respuestas centradas, se realiza el algoritmo para calcular los componentes PLS vía NIPALS.

5. Se hace la regresión por mínimos cuadrados de Y estandarizada con respecto a los componentes PLS obtenidos anteriormente.

6. Se regresa analíticamente a la forma Y contra X^o, es decir, $Y = X\beta^*$, para obtener los coeficientes β^*.

De lo descrito en esta tesis de forma general es importante notar que se esta brindando otro enfoque, para cuando se trabaja con datos ordinales. Otro punto es el considerar otros enfoques para el tema de la salud, con respecto al tabaquismo en México. Esto fue posible gracias a la adaptación que se hizo del trabajo del 2012 de Russolillo [14]. Como cualquier otro enfoque ya sea reciente o no, tiene ventajas y desventajas. Las ventajas de la aplicación de este modelo, giran alrededor de no aumentar la dimensionalidad de las variables predictoras, mantener el orden intrínseco de las variables ordinales, mantener la estructura de los modelos lineales con la cual se tiene teoría que los respalda. La desventaja de este modelo es que al aplicar la transformación de Kruskal, ya no se puede regresar a los valores originales. Sin embargo se puede relacionar las variables predictoras con las variables respuesta. Otra desventaja es que no se cuenta con propiedades asintóticas para los coeficientes de regresión PLS, por tal motivo se recurre a la metodología Bootstrap. El trabajo realizado requirió considerablemente de aspectos computacionales y numéricos. Cabe mencionar que los resultados de este trabajo han sido motivados dentro de un contexto de Bioquímica médica, pero también pudieran aplicarse en áreas ajenas a ésta.

Con respecto al marco teórico del ejemplo, el tabaquismo es de las causas prevenibles de muerte más cuantiosas en el mundo, por lo que un gran avance contra esta enfermedad es un avance para salvar muchas vidas. Aunque no existe una relación exacta comprobada entre los niveles plasmáticos que se manejaron y el riesgo a contraer enfermedades pulmonares, el modelo resultó ser sensato cuando predice los valores de los niveles químicos en sangre y orina. Dentro de lo concerniente a las clasificaciones y crear escenarios, a pesar que la entrada del modelo es la base de datos conjunta, implícitamente clasificó mejor a las mujeres y a los fumadores activos, probablemente podamos atribuir estos resultados a cuestiones de hábitos y de metabolización.

Finalmente el trabajo de tesis plantea puntos a los que pueden darse seguimiento. Entre los puntos principales se identificaron los siguientes

- Realizar la metodología antes descrita para una muestra mayor y ver la estabilidad de los coeficientes.
- Considerar en el modelo predictivo la inclusión de variables categóricas, es decir, variables nominales sin ningún orden explícito.
- Considerar más variables clasificadoras para la creación de diferentes escenarios y evaluar el rendimiento de clasificación entre grupos.
- Considerar el enfoque en Cantaluppi [4] y comparar los resultados con lo que se presentaron en este trabajo.

Apéndice A

Cuestionario

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

ENCUESTA PARA LOS PARTICIPANTES DEL PROYECTO ANÁLISIS TOXICOLÓGICOS Y TOXICOGENÓMICOS PARA UN GRUPO SELECCIONADO DE CARCINÓGENOS DEL TABACO EN SUJETOS JÓVENES VOLUNTARIOS.

Estimado alumno de la Universidad Autónoma de Nuevo León, le solicitamos responder el siguiente cuestionario de forma **voluntaria**. Este cuestionario es parte de un estudio que se está realizando en la Facultad de Medicina. La información que nos proporcione es muy valiosa y nos ayudará a tener más datos sobre el hábito de fumar en la comunidad universitaria. Sus respuestas serán mantenidas **en forma estrictamente confidencial y ninguna persona podrá ser identificada** durante y después del estudio.

I. DATOS PERSONALES

Clave: _____Edad: _____ Sexo: _____ Estatura: _____ Raza: _____

Peso:_____ IMC:_____ Ciudad de procedencia: _____

2¿En qué estado de la Republica nacieron sus abuelos?
Abuelo materno: _____
Abuela materna: _____
Abuelo paterno: _____
Abuela paterna: _____

3. Nivel educativo más alto alcanzado:

- o Ninguno
- o Primaria
- o Secundaria
- o Escuela Técnica
- o Preparatoria
- o Licenciatura
- o Posgrado
- o Otro ¿Cuál?:_____

4. Nivel de ingreso familiar mensual (Moneda nacional):

- o Menos de $ 5,000.00
- o $ 5,000.00 a $ 10,000.00
- o $ 10,000.00 a $ 15,000.00
- o $ 15,000.00 a $ 19,999.00
- o $ 20,000.00 a $ 29,999.00
- o $ 40,000.00 a $ 49,999.00
- o $ 50,000.00 a $ 59,999.00
- o Más de $ 60,000.00

Número de personas que dependen de este ingreso: _____

5. Antecedentes familiares

Familiares con cáncer de pulmón:_____

Familiares con cualquier otro tipo de cáncer:_____

6. Antecedentes personales patológicos

Diabetes mellitus: Sí ____ No ____ Hipertensión arterial Sí____ No____

¿Qué enfermedades le han sido diagnosticadas? _____

_____

Algún tipo de cáncer: _____

II. CONSUMO DE TABACO

7. ¿Es usted fumador / a?
a. Sí
b. No

8.- Sí la respuesta de la pregunta 7 es No, ¿Alguna vez fumó?
a. Sí
b. No c. ¿Hace cuantos años? _____

9. ¿Cuántos años tenía cuando fumó por primera vez?
a. Nunca he fumado
b. Edad:

10	11	12	13	14	15
16	17	18	19	20	21
MAS DE 21					

10. ¿Fuma diariamente?
a. Sí
b. No
¿A qué edad comenzó a fumar diariamente?

10	11	12	13	14	15
16	17	18	19	20	21
MAS DE 21					

11. Durante los pasados 15 días, ¿cuántos días fumó?
a. Todos los días
b. Más de la mitad del tiempo 8-15 días
c. Menos de la mitad del tiempo (1-7 días)
d. Solo a veces (1-5 días)
e. Rara vez (1-3 días)
e. Ninguno

12. Durante los pasados 15 días, en los días en que fumó, ¿cuántos cigarros consumió?
a. Más de 20 cigarros por día (más de 1 cajetilla)
b. 11 a 20 cigarros por día (más de media cajetilla)
c. 6 a 10 cigarros por día (más de cuarto de cajetilla)
d. 2 a 5 cigarros por día
e. 1 cigarro por día
f. Menos de un cigarro por día
g. No fumé durante los pasados 30 días (un mes)

13. Durante los pasados 15 días, ¿qué marca de cigarros fumó con mayor frecuencia? (seleccionar una sola respuesta)
a. No fumé cigarros durante los pasados 30 días (un mes)
b. Ninguna marca especial
c. Marlboro
d. Broadway
e. Boots
f. Montana
g. Camel
h. Otra marca, ¿cuál? _____

14. Durante los pasados 15 días (un mes), ¿alguna vez utilizó tabaco en otra forma que no fueran cigarros? (por ejemplo: puros, pipa, cigarros pequeños, etc.)
a. Sí b. ¿Cuál?_____
c. No

15.- Detalle su consumo de cigarrillo lo mejor posible de los últimos 3 días
Anteayer Número de cigarrillos en la mañana_____
Número de cigarrillos en la tarde _____
Número de cigarrillos en la noche _____
Ayer Número de cigarrillos en la mañana_____
Número de cigarrillos en la tarde _____
Número de cigarrillos en la noche _____
Hoy Número de cigarrillos en la mañana_____
Número de cigarrillos en la tarde _____

16.- ¿Cuantas horas lleva sin fumar el día de hoy? ____

III. EXPOSICIÓN AL HUMO DEL TABACO AMBIENTAL AJENO

17. ¿Ha estado expuesto al humo del tabaco las últimas 48 horas?
a. Sí
b. No

18. Alguien fumó en su presencia en lugares abiertos en las últimas 48 horas?
a. Sí
b. No

19. Durante las últimas 48 horas ¿alguien ha fumado en espacios cerrados en su presencia?
a. Sí
b. No

20. ¿Con qué frecuencia alguna persona fuma en lugares cerrados en su presencia?
a. Todos los días
b. Todas las semanas
c. Todos los meses
d. Menos de una vez al año
e. Nunca

21. ¿Alguien fumó dentro de alguno de los bares, clubes nocturnos ó centros recreativos a los que ha ido en las últimas 48 horas?
a. Sí
b. No
c. No he ido a ese tipo de lugares

22. ¿Cuántas personas fumadoras viven en su casa____

23. ¿Con qué frecuencia alguien fuma dentro de su casa?
a. Todos los días
b. Al menos una vez por semana
c. Al menos una vez al mes
d. Al menos una vez al año
e. Nunca

24. Durante los pasados 15 días, ¿utilizó productos que contienen nicotina para dejar de fumar? (por ejemplo: chicles, parches, enjuagues bucales, etc.)
a. Sí b. ¿Cuál?_____
c. No

IV. HÁBITOS ALIMENTICIOS Y SEDENTARISMO

25.- ¿Consume carnes rojas?
a. No
b. Menos de 5 veces (1500g de carne)/semana
c. 5-15 veces (1500-4500 g de carne)/semana
d. Más de 15 veces (4500 g de carne)/semana
Años consumiendo: _____
Tiempo sin consumir: _____

26. ¿Consume comida procesada al carbón o a la leña?
a. Sí
b. No

27. ¿Ha estado expuesto al humo del carbón o leña?
a. No
b. Una vez al mes
c. Los fines de semana
d. Todos los días

28. ¿Consume frutas y verduras diariamente?
a. No
b. Menos de 100 g/día
c. 110-200 g/día
d. Más de 210 g/día

29.- ¿Consume alcohol?
a. No
b. Menos de 5 copas (75 mL de etanol)/semana
c. 5-10 copas (75-150 mL de etanol)/semana
d. Más de 10 copas (150 mL de etanol)/semana
Años tomando: _____
Tiempo sin tomar: _____

30. Tomando en cuenta diversas formas de actividad física, (trote, carrera, bicicleta, natación) ¿usted le dedica más de 90 minutos en total por semana?
a. Sí
b. No

Estas son todas las preguntas
Muchas gracias por su participación

Nombre y firma de quien realizó la encuesta:_____
Fecha de obtención de la muestra: _____ _____ _____
Día Mes Año

Apéndice B

Algoritmos computacionales

B.1. Algoritmo PLSR

Se tienen dos conjuntos de datos de la forma $X_1 = [X_{11}, X_{12}, ..., X_{1P_1}]$ y $X_2 = [X_{21}, X_{22}, ..., X_{2P_2}]$. A continuación se describe el funcionamiento del algoritmo en la Figura 1.1.

Paso I: Para $h = 1$, $E_0 = X_1$, $F_0 = X_2$ y se inicia con un vector de pesos iniciales $w_{1(1)}$.

Se calcula $t_{1(1)} = E_0 w_{1(1)}$, de aquí se obtiene

$$w_{2(1)} = \frac{F_0^T t_{1(1)}}{||F_0^T t_{1(1)}||}.$$

Para poder obtener $t_{2(1)} = F_0 w_{2(1)}$ y entonces actualizar $w_{1(1)}$ con

$$w_1 = \frac{E_0^T t_{2(1)}}{||E_0^T t_{2(1)}||}.$$

Este paso se realiza hasta que se tiene convergencia en $w_{1(1)}$ y con esto se obtiene también la convergencia en $w_{2(1)}$.

Paso II: Se calcula el estimador de la regresión de E_0 sobre $t_{1(1)}$

$$P_{(1)} = \frac{E_0^T t_{1(1)}}{t_{1(1)}^T t_{1(1)}}. \tag{B.1}$$

Paso III: Se calculamos el coeficiente de regresión de $t_{2(1)}$ sobre $t_{1(1)}$

$$b_{(t_{2(1)}|t_{1(1)})} = \frac{t_{2(1)}^T t_{1(1)}}{t_{1(1)}^T t_{1(1)}}. \tag{B.2}$$

Paso IV: Se extraen los residuos de la regresión del **Paso II**

$$E_1 = E_0 - t_{1(1)} P_{(h)}^T.$$

Paso V: Se obtienen los residuos de la regresión del **Paso III**

$$F_1 = F_0 - b_{(t_{2(1)}|t_{1(1)})} t_{1(1)} w_{2(1)}^T.$$

Paso VI: Se repiten los **Pasos I-V** hasta obtener el número de componentes PLS deseados.

B.2. Algoritmo NMPLSR

Se tienen dos conjuntos de datos de la forma $X_1 = [X_{11}, X_{12}, ..., X_{1P_1}]$ y $X_2 = [X_{21}, X_{22}, ..., X_{2P_2}]$. A continuación se describe el funcionamiento del algoritmo en la Figura 2.1.

Paso I: Para $h = 1$ se inicia con un $t_{2(1)}$

Para toda $p = 1, ..., P_1$ se realiza lo siguiente

$$\hat{x}_{p1} \propto \mathbf{Q}(\tilde{X}_{p1}, t_{2(1)}).$$

Se calculan los pesos de $\hat{X}_2$ como

$$w_{1(1)} = \frac{\hat{X}_1^T t_{2(1)}}{||\hat{X}_1^T t_{2(1)}||}.$$

Se obtiene la variable latente

$$t_{1(1)} = \hat{X}_1 w_{1(1)}.$$

Para toda $p = 1, ..., P_2$ se obtiene

$$\hat{x}_{p2} \propto \mathbf{Q}(\tilde{X}_{p2}, t_{1(1)}).$$

Ahora se calculan los pesos de $\hat{X}_1$

$$w_{2(1)} = \frac{\hat{X}_2^T t_{1(1)}}{||\hat{X}_2^T t_{1(1)}||}.$$

Se actualiza el valor de $t_{2(1)}$ como

$$t_{2(1)} = \hat{X}_2 w_{2(1)}.$$

Se repite el Paso I hasta que se tenga la convergencia en $w_{1(1)}$ y con esto se obtiene también la convergencia en $w_{2(1)}$.

Se realiza el algoritmo de la Figura 1.1 para $h = 2, ..., H$.

Paso II: Para $h = 2$, se inicia con un vector de pesos iniciales $w_{1(2)}$.

Se calcula $t_{1(2)} = E_1 w_{1(2)}$, de aquí se obtiene

$$w_{2(2)} = \frac{F_0^T t_{1(2)}}{||F_0^T t_{1(2)}||}.$$

Para poder obtener $t_{2(2)} = F_1 w_{2(2)}$ y entonces se actualiza $w_{1(2)}$ con

$$w_1 = \frac{E_1^T t_{2(2)}}{||E_0^T t_{2(2)}||}.$$

Este paso se realiza hasta que se tenga convergencia en $w_{1(2)}$ y con esto se obtiene también la convergencia en $w_{2(2)}$.

Paso III: Se calcula el estimador de la regresión de E_0 sobre $t_{1(2)}$

$$P_{(2)} = \frac{E_0^T t_{1(2)}}{t_{1(2)}^T t_{1(2)}}. \tag{B.3}$$

Paso IV: Se calcula el coeficiente de regresión de $t_{2(2)}$ sobre $t_{1(2)}$

$$b_{(t_{2(2)}|t_{1(2)})} = \frac{t_{2(2)}^T t_{1(2)}}{t_{1(2)}^T t_{1(2)}}. \tag{B.4}$$

Paso V: Se extraen los residuos de la regresión del **Paso II**

$$E_1 = E_0 - t_{1(2)} P_{(h)}^T.$$

Paso VI: Se obtienen los residuos de la regresión del **Paso III**

$$F_1 = F_0 - b_{(t_{2(2)}|t_{1(2)})} t_{1(2)} w_{2(2)}^T.$$

Paso VII: Se repiten los **Pasos II-VI** hasta obtener el número de componentes PLS deseados (H).

Apéndice C

T. Monótona de Mínimos Cuadrados de Kruskal

```
dummy.ord<-function (Y,X)
{
X<-as.matrix(X)       # Matriz a transformar
Y<-as.matrix(Y)       # Vector que brinda un orden a X
n<-nrow(X);p<-ncol(X) # Dimensiones de X
Ypred<-matrix(0,n,p)  # Matriz nula para $X^o$
eta2<-array(,p)       # Vector nulo, de longitud igual a las columnas de X
for (k in 1:p) {      # Índice que recorre las p columnas de X
Xtemp<-matrix(0,n,length(levels(as.factor(X[,k]))))
for (j in 1:length(levels(as.factor(X[,k])))) {
for (i in 1:length(X[,k])) {
if (is.na(X[i,k])) {
Xtemp[i,j] = NA
}
else {
if (X[i,k]==j) {
Xtemp[i,j] = 1
}}}}
Q<-as.vector(tapply(Y,X[,k],mean,na.rm=T))
#print(Q)
repeat {
ncol_XtempOld<-ncol(Xtemp)
for (i in 1:(ncol(Xtemp)-1)) {
#print(paste("i=",i))
if (Q[i]>Q[i+1]) {
Xtemp[,i+1]<-Xtemp[,i]+Xtemp[,i+1]
Xtemp<-as.matrix(Xtemp[,-i])
```

```
#print(Xtemp)
if (ncol(Xtemp)==1) {
# print("Se mantuvo una columna")
Q <- c()
for (i in 1:ncol(Xtemp)) {
Q[i] <- sum((Xtemp[,i])*Y,na.rm=T)/sum(Xtemp[,i], na.rm=T)
}
}
else {
Q <- c()
for (i in 1:ncol(Xtemp)) {
Q[i] <- sum((Xtemp[,i])*Y,na.rm=T)/sum(Xtemp[,i], na.rm=T)
}
}
break}}
ncol_XtempNew<-ncol(Xtemp)
if (ncol(Xtemp)==1) {break}
else {
#print("ncol_XtempOld")
#print(ncol_XtempOld)
if (ncol_XtempNew==ncol_XtempOld) {break}
}
}
Ypred[,k]<-(Xtemp%*%Q)
eta2[k]<-var(Ypred[,k],na.rm = T)/var(Y, na.rm = T)
#Si quiere ponderar una variable dividiendola por la raíz del numero de grupos:
#Ypred[,k]<-(Xtemp%*%(as.vector(tapply(Y,X[,k],mean,na.rm=T))))/
length(levels(as.factor(X[,k])))^(1/2)
if (k==1) {
Xdummy<-Xtemp
}
else {
Xdummy<-cbind(Xdummy,Xtemp)
}}
list(Xdummy=Xdummy,Quant=Ypred, eta2=eta2, Q=Q) # Se almacenan los resultados paramét
}
```

Apéndice D

Códigos en R

D.1. Código para NMPLS vía NIPALS

```
N1<-dim(X1)[1]                # Número de individuos (99)
P1<-dim(X1)[2];P2<-dim(X2)[2]# Preguntas predictoras y Variables de respuesta
H<-P1                         # Núm. de componentes PLS (TODOS LOS QUE SE PUEDAN)
B<-c()                        # Vector que guarda los coef OLS de t_2 vs t_1
rmse<-rep(0,H)                # Vector que guarda el Root Square Medium Error
matrizP<-matrix(0, P1, H)     # Matriz para acomodar los coef OLS de X_1 vs t1
hatX1<-matrix(0, N1, P1)      # Matriz para acomodar los valores de \hat(X1)
hatx2<-matrix(0, N1, P2)      # Matriz para acomodar los valores de \hat(X2)
matrizw1<-matrix(0, P1, H)    # Matriz para acomodar los w_1(h)
matrizw2<-matrix(0, P2, H)    # Matriz para acomodar los w_2(h)
matrizt1<-matrix(0, N1, H)    # Matriz que guarda los componentes PLS de t_1
matrizt2<-matrix(0, N1, H)    # Matriz que guarda los componentes PLS de t_2
w1.1<-rep(1,P1)               # Se inicia con pesos
w1.1old<-rep(0,P1)            # vector de 0's

#Se empieza el algoritmo para h=1
cont<-rep(0, H)     # Vector que guarda el numero de iteraciones en cada ciclo while
while(sum(abs(w1.1old-w1.1))>0.000000000000001){
  w1.1old<-w1.1               # Se guarda el valor de w1.1
  t1.1<-X1%*%w1.1             # Se calcula t_1(1)
  for(i in 1:P2){             # Se actualiza X2
    hatx2[,i]<-Qadhoc(x2[,i],t1.1)
    }
  aux1<-as.vector(t(hatx2)%*%t1.1)
  w2.1<-(aux1)*(1/norma(aux1))
  t2.1<-hatx2%*%w2.1
```

```
  for(i in 1:2){             # Las primeras dos variables no se transforman
    hatX1[,i]<-X1[,i]        # ya que son continuas
  }
  for(i in 3:P1){         # Se actualiza X1
      hatX1[,i]<-Qadhoc(X1[,i],t2.1)
    }
  aux2<-as.vector(t(hatX1)%*%t2.1)
  w1.1<-(aux2)*(1/norma(aux2))
  cont[1]<-cont[1]+1
}
  matp1<-as.vector(t(hatX1)%*%t1.1)/(t(t1.1)%*%t1.1)
   b2.1<-as.numeric((t(t2.1)%*%t1.1)/(t(t1.1)%*%t1.1)[1,1])
     E1<-hatX1-t1.1%*%matp1
rmse[1]<-sqrt(sum(E1*E1))
     F1<-hatx2-(b2.1*t1.1)%*%w2.1
      B<-c(B,b2.1)
matrizw1[,1]<-w1.1;matrizw2[,1]<-w2.1
matrizt1[,1]<-t1.1;matrizt2[,1]<-t2.1
 matrizP[,1]<-matp1

for(h in 2:H){
     w1.h<-rep(1/P1,P1)
  w1.hold<-rep(0,P1)
  while(sum(abs(w1.h-w1.hold))>0.00000000000000000001){
   w1.hold<-w1.h           # Se guarda el valor de w1.1
   t1.h<-as.vector(E1%*%w1.h)
   w2.h<-as.vector(t(F1)%*%t1.h)*(1/norma(t(F1)%*%t1.h))
   t2.h<-as.vector(F1%*%w2.h)
   w1.h<-as.vector(t(E1)%*%t2.h)*(1/norma(t(E1)%*%t2.h))
   cont[h]<-cont[h]+1
  }
  matph<-as.vector(t(E1)%*%t1.h)/(t(t1.h)%*%t1.h)
   b2.h<-as.numeric((t(t2.h)%*%t1.h)/(t(t1.h)%*%t1.h)[1,1])
     E1<-E1-t1.h%*%t(matph)
rmse[h]<-sqrt(sum(E1*E1))
     F1<-F1-(b2.h*t1.h)%*%t(w2.h)
      B<-c(B,b2.h)
  matrizw1[,h]<-w1.h;matrizw2[,h]<-w2.h
  matrizt1[,h]<-t1.h;matrizt2[,h]<-t2.h
   matrizP[,h]<-matph
}
```

D.1.1. Funciones auxiliares

Las funciones *Qadhoc* y *norma* se programaron de la siguiente forma.

```
Qadhoc<-function(a,b){     # Calcula la matriz de proyección
  aux<-solve(t(a)%*%a)     # Q = a(a^T a)-1 a^T b
  return(a%*%aux%*%a%*%b)
  }

norma<-function(x){        # Calcula la norma de X
  return(sqrt(t(x)%*%x))   # ||X||=(x_1^2+...+x_n^2)^1/2
  }
```

D.2. Código para el análisis de regresión

El siguiente código es para recuperar la forma $Y = X\beta^*$ para cada componente PLS

```
h<-8                 # Se fija el número de componentes a transformar, es decir,
vec.bts<-c();vec.bts # se establece el número de componentes PLS a considerar.
for(j in 1:P2){
  vectnull<-c()
  vectnull<-c(vectnull,matbets[1,j]*matrizw1[,1])
  for(k in 2:h){
    auxmul<-diag(P1)
    for(i in 1:(k-1)){
      auxx<-diag(P1)-(matrizw1[,i]%*%t(matrizP[,i]))
      auxmul<-auxmul%*%auxx
    }
    aaux<-matbets[k,j]*auxmul%*%matrizw1[,k]
    vectnull<-vectnull+aaux
  }
  vec.bts<-cbind(vec.bts, vectnull)
}
BTSh<-vec.bts;BTSh
```

D.3. Intervalos por percentiles

El siguiente código es para obtener los intervalos Bootstrap para los coeficientes de regresión

```
X.Boot<-BASEPLS # Es la X^* de donde se va a muestrear con reemplazo
Y.Boot<-YbC     # Y's centradas de donde se va muestrear con reemplazo
> dim(X.Boot)
[1] 99  8       # 99 sujetos y 8 predictores

# Se combinan los predictores con cada variable respuesta,
# deben de permanecer juntos para cuando se haga el remuestreo
# se tomen los biomarcadores como debe ser (respecto a su índice).
Ind1<-cbind(XX.Boot,Y.Boot[,1])
Ind2<-cbind(XX.Boot,Y.Boot[,2])
Ind3<-cbind(XX.Boot,Y.Boot[,3])
Ind4<-cbind(XX.Boot,Y.Boot[,4])

# NMPLSR.BOOT es la función de donde se calcula el estadístico de interés
# este tiene una entrada ''index'' el cual dicta el muestreo con reemplazo
NMPLSR.BOOT<-function(X, index){
  Xb<-X[index,-9]              # Se extraen los predictores
  Yb<-X[index,-1:-8]           # Se extrae el vector respuesta
  N1<-dim(Xb)[1]               # Número de individuos
  P1<-dim(Xb)[2];P2<-1         # Preguntas predictoras y Variables de respuesta
  H<-P1                        # Num de componentes PLS (TODOS LOS QUE SE PUEDAN)
  B<-c()                       # Vector que guarda los coef OLS t2 vs t1
  matrizP<-matrix(0, P1, H)    # Matriz para acomodar los p(h)
  hatXb<-matrix(0, N1, P1)     # Matriz para acomodar los valores de \hat(Xb)
  hatYb<-matrix(0, N1, P2)     # Matriz para acomodar los valores de \hat(Yb)
  matrizw1<-matrix(0, P1, H)   # Matriz para acomodar los w_1(h)
  matrizw2<-matrix(0, P2, H)   # Matriz para acomodar los w_2(h)
  matrizt1<-matrix(0, N1, H)   # Matriz que guarda los componentes PLS de t_1
  matrizt2<-matrix(0, N1, H)   # Matriz que guarda los componentes PLS de t_2
  w1.1<-rep(1,P1)
  w1.1old<-rep(0,P1)
  cont<-rep(0, H)
  while(sum(abs(w1.1old-w1.1))>0.000000000000001){
    w1.1old<-w1.1
    t1.1<-Xb%*%w1.1
    hatYb<-Qadhoc(Yb,t1.1)
    aux1<-as.vector(t(hatYb)%*%t1.1)
    w2.1<-(aux1)*(1/norma(aux1))
```

```
    t2.1<-hatYb%*%w2.1
    for(i in 1:2) {hatXb[,i]<-Xb[,i]}
    for(i in 3:P1){hatXb[,i]<-Qadhoc(Xb[,i],t2.1)}
    aux2<-as.vector(t(hatXb)%*%t2.1)
    w1.1<-(aux2)*(1/norma(aux2))
    cont[1]<-cont[1]+1
  }# Como ya convergió w1.1 calculamos lo del siguiente ciclo
  matp1<-as.vector(t(hatXb)%*%t1.1)/(t(t1.1)%*%t1.1)
  b2.1<-as.numeric((t(t2.1)%*%t1.1)/(t(t1.1)%*%t1.1)[1,1])
  E1<-hatXb-t1.1%*%matp1
  F1<-hatYb-(b2.1*t1.1)%*%w2.1
  matrizw1[,1]<-w1.1;matrizw2[,1]<-w2.1
  matrizt1[,1]<-t1.1;matrizt2[,1]<-t2.1
  matrizP[,1]<-matp1;           B<-c(B,b2.1)
  for(h in 2:H){
    w1.h<-rep(1/P1, P1)
    w1.hold<-rep(0, P1)
    while(sum(abs(w1.h-w1.hold))>0.00000000000000001){
      w1.hold<-w1.h           # Se guarda el valor de w1.1
      t1.h<-as.vector(E1%*%w1.h)
      w2.h<-as.vector(t(F1)%*%t1.h)*(1/norma(t(F1)%*%t1.h))
      t2.h<-as.vector(F1%*%w2.h)
      w1.h<-as.vector(t(E1)%*%t2.h)*(1/norma(t(E1)%*%t2.h))
      cont[h]<-cont[h]+1
    }
    matph<-as.vector(t(E1)%*%t1.h)/(t(t1.h)%*%t1.h)
    b2.h<-as.numeric((t(t2.h)%*%t1.h)/(t(t1.h)%*%t1.h)[1,1])
    E1<-E1-t1.h%*%t(matph)
    F1<-F1-(b2.h*t1.h)%*%t(w2.h)
    matrizw1[,h]<-w1.h;matrizw2[,h]<-w2.h
    matrizt1[,h]<-t1.h;matrizt2[,h]<-t2.h
    matrizP[,h]<-matph
    B<-c(B,b2.h)
  }

############### SE HACE LA REGRESIÓN ################################

  t.s<-matrizt1              # Se almacena la matriz de componentes PLS
  beta1<-lm(Yb~t.s)$coef[-1] # Se elimina el intercepto
matbets<-cbind(beta1)        # Se almacenan los coef. de la regresión

# Se recupera la forma original #
```

```
  vec.bts<-c();vectnull<-c() # vector para las betas y uno auxiliar
  vectnull<-c(vectnull,matbets[1]*matrizw1[,1]) # se hace para h=1 en
  for(k in 2:8){ # Se establece el número de componentes PLS a considerar (8)
    auxmul<-diag(P1)
    for(i in 1:(k-1)){
      auxx<-diag(P1)-(matrizw1[,i]%*%t(matrizP[,i]))
      auxmul<-auxmul%*%auxx
    }
    aaux<-matbets[k]*auxmul%*%matrizw1[,k]
    vectnull<-vectnull+aaux
  }
  vec.bts<-cbind(vec.bts, vectnull)
  } # fin del for(k)
  colnames(vec.bts)<-c("Indicador")
  rownames(vec.bts)<-c("Edad","IMC","V3","V4","V5","V6","V7","V8")
  BeTaS<-vec.bts
  return(BeTaS)
  }
# Fin de la función NMPLSR.BOOT
# Se hacen las réplicas Bootstrap con la función ‘‘boot’’,
# la cual está en la paquetería del mismo nombre.
```

Bibliografía

[1] Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. 2da Ed, John-Wiley & Sons.

[2] Agresti, A. (2002). *Categorical Data Analysis*. 2da Ed, John-Wiley & Sons.

[3] Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

[4] Cantaluppi, G. (2012). "A Partial Least Squares Algorithm Handling Ordinal Variables Also In Presence Of A Small Number Of Categories".

[5] Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

[6] Garthwaite, P. (1994). "Interpretation of Partial Least Squares". *Journal of the American Statistical Association*, Vol. 89, No. 425, pp. 122-127.

[7] Hoskuldsson, A. (1988). "PLS regression Methods". *Journal of Chemometrics*, Vol. 2. pp. 211-228.

[8] Ildiko, F. & Friedman, J. (1993). "A Statistical View of Some Chemometrics Regression Tools". *American Society for Quality*, Vol. 35. pp. 109-135.

[9] Johnson, R & Wichern, D. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

[10] Kruskal, J. (1964). "Nonmetric Multidimensional Scaling: A Numerical Method". *Psychometrika*, Vol. 29, No. 2, pp. 115-129.

[11] Kruskal, J. (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". *Psychometrika*, Vol. 29, No. 1, pp. 1-27.

[12] Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw-Hill.

[13] R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Disponible en: http://www.r-project.org.

[14] Russolillo, G. (2012). "Non-Metric Partial Least Squares". *Electronic Journal of Statistics*, Vol. 6, pp. 1641 - 1669.

[15] Russolillo, G. & Natale-Lauro C. (2011). "A Proposal for Handling Categorical Predictors in PLS Regression Framework".

[16] SAS Institute (2013). *SAS Base version 9.3*, Cary Carolina del Norte, E.U.A.

[17] Skrondal, A & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. Chapman & Hall.