Universidad de Concepción
Dirección de Postgrado
Facultad de Ciencias Físicas y Matemáticas
Programa de Magíster en Estadística

Métodos de Regularización para la Selección de Variables Aplicados a la Predicción del Riesgo de Padecer disfunción motora en Adultas Mayores Activas de la Ciudad de Valdivia.

Tesis Para Optar al Grado de Magíster en Estadística

MARIA JOSE ALEJANDRA MEDINA FRITZ
CONCEPCIÓN-CHILE
2018

Profesor Guía: María Paz Casanova Laudien
Departamento de Estadística
Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción

UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

TESIS PARA OPTAR AL GRADO DE MAGISTER EN ESTADISTICA

Métodos de Regularización para la Selección de Variables Aplicados a la Predicción del Riesgo de Padecer Disfunción Motora en Adultas Mayores Activas de la Ciudad de Valdivia.

Profesor Guía	:	María Paz Casanova Laudien	Firma	.....
Profesor Colaborador	:	Nora Serdyukova	Firma	.....
Profesor Consejero	:	Sebastián Andrés Niklitschek Soto	Firma	.....
Nombre Memorante	:	María José Alejandra Medina Fritz	Firma	.....
Teléfono	:	(9)96764051		
e-mail	:	mmedinaf@udec.cl		

Concepción, 2018

Dedicado a
mis hijos

Agradecimientos

A Dios.
Por estar a mi lado, ayudandome siempre a lograr mis objetivos, por protegerme a mi y a los que amo, por guiar mi camino y entregarme siempre lo mejor.

A mi familia.
Gracias viejitas lindas (Leonor y Gladys) por creer en mi y apoyarme en este nuevo sueño. Zasha por acompañarnos en estos últimos meses y Julietita por despertarme para trabajar, por ser tan linda y tiernita.

A mis hermanos Leonardo, Eduardo y Gabriela.
Por estar pendiente de mi, preocuparse por cada paso que doy, quererme y pretegerme en todo momento.

A mis profesores.
Doctora María Paz Casanova Laudien, agradezco sinceramente que haya creído en mi, que me haya apoyado cuando más la necesitaba. Además, quiero agradecer todo el tiempo que me entregó, incluso en parte de sus vacaciones, su confianza, cariño y dedicación.
También quiero dar gracias a la profesora Nora Serdyukova por todas sus gestiones y por querer ser parte de este trabajo junto al profesor Sebastián Niklitschek quien tuvo la disposición de estar nuevamente aportando en este periodo de finalización.

A mis amigas y amigos
Catalina Ávila, Denisse Vidal, Karen Araneda, María Paz Tapia, Nataly Zapata, aunque ya no estamos juntitas desde lejos me siguen apoyando y enviando todas sus buenas vibras. Gracias por haber sido y seguir siendo parte

de mi vida. Tambien quiero agradecer al Francisco Toledo (Perrin), por haberme ayudado cuando peor me sentía, cuando realmente necesite un amigo tu estabas ahí, gracias perrin por todo tu apoyo. Por último a christian Nuñez por su buena onda y sus buenas vibras, siempre estar para entregar alegría y apoyar en lo que nos proponemos.

Pero por sobre todo, quiero agradecer a quienes están siempre alentandome, escuchandome, dandome su amor y dedicación.

A mis hijos Matías e Ignacia.
Quienes siguen siendo el pilar fundamental de todo lo que hago y me propongo. Nuevamente entregaron sus tiempos para que yo pudiera lograr este proceso. Agradezco también su apoyo, energía, comprensión y cariño, lo cual en muchas oportunidades me ayudó a seguir.

En general les agradezco por existir, por ser tal como son, sin ustedes bebés no podría seguir ni hubiese logrado tanto. Me llena de orgullo saber que son y seguirán siendo unas grandes personas llenas de amor, fuerza y entrega. Los amodoro!!
***Matías**; gracias bebé por ser mi hijo, por entenderme siempre y por tu apoyo. Te estás convirtiendo en un gran hombre, me siento muy orgullosa de ti. Gracias por ser tú.*
***Ignacia**; pequeñita muchas gracias por tus abracitos, ojalá que nunca se acaben, me llenan de amor; gracias por tu alegría y desplante me ayudas a ver la vida con optimismo.*

A mi Pablo.
Por hacerme inmensamente feliz, entregarme todo su amor, comprensión, por esas lindas palabras y también por escucharme y apapacharme cuando más lo necesité.
Amor, gracias por toda tu dulzura, por creer en mi, decirme siempre que todo lo puedo, que te sientes orgulloso de mi y también por darte la lata de leer el artículo. Todas tus palabras, tu amor y comprensión me han ayudado durante esta étapa.

Quiero agradecer en general a la vida, por todo lo que me entrega, por todos ustedes, bellas personas que ha puesto en mi camino. Lo tengo todo, sólo queda agradecer. ¡¡ Gracias!!

Resumen

El presente trabajo aborda el problema de la disfunción motora (limitación de la capacidad de movimiento que provoca una disminución del rendimiento o restricción en la ejecución de funciones o acciones motoras consideradas normales) en adultas mayores pertenecientes a la ciudad de Valdivia. Los métodos estadísticos clásicos no son adecuados para resolver este problema en particular, porque, en este caso existe una alta cantidad de variables, por lo tanto, la implementación y adaptación de los métodos estadísticos modernos como lo son LASSO y Elastic Net (E-N) son primordiales para la solución de tal problema.

En Chile, las tasas de Obesidad y Sedentarismo son muy elevadas. Por una parte, la Obesidad en mujeres aumentó desde 30,7 % en 2010 a 38,4 % en 2017. Este patrón se comienza a observar en niños pre escolares, en efecto, actualmente el 11 % presenta esta condición. Por otra parte, el 86,7 % de la población es sedentaria. Estos dos factores están asociados a disfunción motora. Otro factor asociado es la Dinapenia, que junto a la Obesidad generan un mayor grado de severidad en la disfunción motora. Se ha demostrado que las personas que presentan tanto Obesidad como Dinapenia tienen un ritmo de caminata más lenta que los que pertenecen a otros grupos.

La Edad, ser mujer y no practicar algún tipo de actividad física también son factores de riesgo que influyen en el deterioro funcional y la pérdida de masa muscular.

La relevancia de éste estudio se explica por la presencia de estos factores de riesgo en un gran porcentaje de las mujeres adultas mayores en Chile.

Esta investigación busca proporcionar un método sencillo de diagnóstico del riesgo de padecer disfunción motora en adultas mayores activas.

En el año 2012 la Universidad Austral de Chile, evaluó a 96 adultas mayores, de las cuales algunas eran activas, y se midió a cada una de ellas 93 variables, considerando como variable de interés la disfunción motora, medida

a través del Test de Marcha 6 Minutos. Se clasificó a las participantes como sanas o según Severidad de la Disfunción (grados 1 a 4), con la finalidad de explorar una forma de predicción del Grado de Disfunción en términos de las variables medidas. Se utilizaron métodos de selección de variables, cuyos resultados se evaluaron en términos de sensibilidad, especificidad y porcentaje de correcta selección.

El método de selección de variables que presentó mejor poder predictivo fue LASSO, reduciendo la cantidad de variables de 93 a 5. Los factores incluidos en el modelo son Perímetro de Cintura, Fuerza de Prensión Manual, Frecuencia Cardiaca en el 4° Minuto, Diabetes Mellitus y clasificación como Obesas Dinapénicas. Este modelo simplifica la evaluación de disfunción motora, pues requiere sólo variables biomédicas de las pacientes y la Frecuancia Cardiaca en el Cuarto Minuto, luego de una marcha sostenida.

Las personas que padecen Diabetes Mellitus, son Obesas Dinapénicas, tienen un Perímetro de Cintura alto, presentan una Frecuencia Cardiaca baja, y además su Fuerza de Prensión Manual es baja, tienen mayor riesgo de padecer disfunción motora.

Tabla de contenido

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Actualmente en Chile, más del 60 % de la población presenta exceso de peso y la prevalencia de Obesidad en adultos aumenta en el tiempo. Según MINSAL (2010), el 30,7 % de las mujeres adultas presentaban algún grado de Obesidad, en 2017 la cifra aumentó a 38,4 % (MINSAL, 2017). Este patrón se comienza a observar en niños pre escolares; en efecto, actualmente el 11 % de ellos presenta esta condición.

Sumado a lo anterior, el 86,7 % de la población es Sedentaria, esta cifra no disminuye según grupos de la población. Para el caso específico de las mujeres, el 90 % de ellas no realiza actividad física de forma frecuente. Afortunadamente las últimas mediciones reflejan una leve disminución respecto de 2010 (MINSAL, 2010, 2017).

Melzer et al. (2004) señalan que la Edad incrementa el deterioro funcional y la pérdida de masa muscular, lo que se agrava más en el caso de las mujeres que no practican ninguna actividad física. En el estudio realizado por Jenkins (2004), se evidenció que la mayor parte de los adultos mayores obesos experimentaron deterioro funcional y se pudo determinar que existe una relación inversa estadísticamente significativa entre el ejercicio y el deterioro funcional.

La Obesidad y los malos hábitos constituyen importantes factores de riesgo para las enfermedades crónicas, como por ejemplo la Diabetes Mellitus (DMII), la cual, según Muñoz (2016), está relacionada con la velocidad motora.

En la conferencia de consenso convocada por la Society on Sarcopenia, Cachexia and Wasting Disorders en Roma 2009, se concluyó que "la Sarcopenia, es decir, la pérdida de masa muscular asociada a la Edad y su prevalencia aumenta con el envejecimiento". En dicha conferencia se definió la Sarcopenia con Movilidad Limitada en adultos mayores "como el hecho de caminar menos de 400 metros durante 6 minutos", y presentar pérdida significativa de masa muscular en comparación a un grupo de personas con edades entre 20 y 30 años. "Clínicamente, una disminución significativa en Sarcopenia se define como un aumento de al menos 50 metros en la caminata de 6 minutos o un aumento de la velocidad al caminar de al menos 0.1 m\s" (Fielding et al., 2011; Morley et al., 2011).

La Dinapenia, por otro parte, se encuentra asociada a la pérdida de fuerza a lo largo de los años, la cual suele evidenciarse, a través de la Fuerza de Prensión Manual (FPM), la cual se utiliza como un predictor de riesgo de discapacidad o disfunción; el criterio o punto de corte se basa en la fuerza de un grupo de referencia y se calcula como el promedio menos dos desviaciones estándar, siguiendo las especificaciones del consenso internacional realizado por la Society on Sarcopenia (Morley et al., 2011). Navia et al. (2012) concluyeron en su estudio que "existen otros factores además de la masa muscular que regulan la pérdida de fuerza muscular del adulto mayor (Dinapenia), por lo tanto, la definición de Sarcopenia debería limitarse sólo a la pérdida de masa muscular relacionada al envejecimiento" (Bouchard et al., 2009). Es así como, tanto la Obesidad como la Dinapenia explican la disfunción motora (DM), factores que en conjunto causan una velocidad de marcha menor que las de los grupo que tienen sólo Obesidad o Dinapenia (Bouchard et al., 2009). Además, en este mismo estudio se concluye que el resultado en el Test de Marcha Seis Minutos (TM6M) fue menor en el grupo con las dos características (Obesidad y Dinapenia) y este mismo grupo presentó una mayor proporción de discapacidad de levantamiento. En la actualidad no existen muchos estudios que expliquen cuáles son las variables que realmente están asociadas a la DM.

Todos estos antecedentes motivan el presente estudio que tiene como objetivos, por un lado identificar cuáles son las variables máyormente relacionadas a esta enfermedad y, por otro lado, proponer un instrumento diagnóstico de fácil utilización.

Capítulo 2

Marco Teórico

2.1. Regresión

En este primer capítulo se dará a conocer en que consiste el modelo de regresión lineal múltiple. Además se introducirán los modelos de regresión penalizada entre estos regresión Ridge, regresión LASSO, y regresión Elastic-Net junto a la generalización de todos estos métodos, regresión Bridge.

2.1.1. Regresión Lineal Múltiple

Como se dijo anteriormente, en muchas ocasiones el comportamiento de una variable respuesta Y no puede ser explicado por una sola variable predictora. En efecto, el modelo es evaluado a través del coeficiente de determinación R^2; cuando el valor de este es bajo, el modelo no es bueno. En este caso se pueden utilizar técnicas para mejorar el modelo. Entre las más utilizadas está la transformación de variables respuesta, independientes o ambas; la utilización de algún método de regresión no paramétrico o agregar más variables al modelo.(Acuña, 2008).

El modelo de regresión múltiple con p variables independientes, queda definido de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

En forma matricial esto se denota por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \overset{iid}{\sim} N_n.$$

donde $\mathbf{Y}$ es un vector columna n dimensional que representa a la variable respuesta, $\mathbf{X}$ es una matriz $n \times p$, para $p = p+1$, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión que será estimado, $\boldsymbol{\epsilon}$ es un vector aleatorio n dimensional, tal que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I_n})$.

2.1.2. Selección de Variables

Como ya se mencionó, en muchas ocaciones la cantidad de variables existentes es demasiado grande y se hace necesario conocer cuales son las variables que realmente son útiles.

Por lo general, a medida que se incorporan variables al modelo, el ajuste de los datos va aumentando, produciendo sobreajuste.

Para poder estimar el vector de parámetros $\boldsymbol{\beta}$ es común utilizar mínimos cuadrados ordinarios. No obstante, este método no siempre es útil, especialmente cuando la cantidad de variables es grande.

Algoritmos de Estimación

Cuando se habla de selección de variables, puede ser de gran utilidad saber cuál es el pensamiento del experto en el área respecto de las variables más importantes, pero esto podría llevar a un error en la correcta selección de las variables (por ejemplo, se podrían incluir variables redundantes, generando multicolinealidad). Es por esto que para poder seleccionar las variables correctas, en número y calidad, es necesario utilizar técnicas de selección. Existen 2 criterios importantes de considerar. En primer lugar, seleccionar las variables que proporcionan mayor información al modelo, y en segundo lugar lograr un número no tan alto de variables seleccionadas, que a la vez entregue la mayor cantidad de información posible y además, sea de fácil interpretación. Desde aquí nace el concepto de "Parsimonia", que se interpreta como un "equilibrio entre la simplicidad (menor número de variables) y el ajuste (tantos regresores como sea necesario)"(NCSS, 2018).

Se llama modelos de regularización para la selección de variables a un amplio conjunto de técnicas de regresión. En la actualidad existen muchos métodos de regularización, algunos se abocan a conjuntos pequeños de variables y otros a conjuntos grandes de variables.

2.2. Stepwise

Los métodos Stepwise o paso a paso, son procedimientos recursivos para la selección de variables, y existen al menos tres tipos que se describen a continuación:

- Fordward: Este método es muy utilizado cuando se tiene un grupo de variables de gran tamaño, desde el cual selecciona generalmente entre 10 y 15 variables. Luego de aplicado el método, este subconjunto puede volver a ser analizado por medio de un método de regresión tradicional, con el objetivo de observar la significancia de estas variables en el modelo. La selección de variables por medio de Forward comienza con un modelo sencillo, que no contiene ninguna variable explicativa, para el cual la primera variable en ingresar será la que presente un mayor coeficiente de correlación en valor absoluto con la variable dependiente. En los pasos siguientes solamente ingresaran variables, la elección se basará en la mejora del R^2. Este procedimiento termina cuando el valor del coeficiente de determinación ya no tiene un aumento significativo (NCSS, 2018).

- Backward: Este método, a diferencia del mencionado anteriormente, comienza con todas las variables disponibles (modelo saturado). Es por esto que el valor del coeficiente de determinación en un comienzo es muy alto. La eliminación de las variables se va realizando de una en una. En un principio se busca cuál es la variable menos significativa, la que será desechada; este procedimiento se realiza hasta que las variables en conjunto proporcionen el mejor ajuste. El mayor problema que presenta esta técnica, es que dentro del subconjunto de selección final pueden quedar variables que no son significativas en el modelo de predicción. Cabe destacar que el nivel de significancia para poder eliminar las variables es propuesto por el investigador (NCSS, 2018).

- Método Stepwise: Fue introducido por Efroymson en el año 1960 con el objetivo de mejorar los dos métodos mencionados anteriormente. Este

método de selección es una combinación del método Forward y el método Backward. En un principio se comienza con el modelo más sencillo, ingresando la variable que en valor absoluto, se encuentra mayormente correlacionada con la respuesta o que entrega un mayor valor de R^2. Luego se analiza la significancia de todas las variables restantes junto con la que ya fue seleccionada, y se verifica si el ajuste del modelo aumenta. En caso de no ser así, la variable es eliminada, en caso contrario la variable es ingresada. El procedimiento continúa intentando ingresar o eliminar variables en cada paso. Este método, necesita de dos niveles de significancia, uno para ingresar variables, y el otro para verificar la mantención de estas en el modelo. Es ideal que el valor de significancia de entrada sea menor que el de salida, con el objetivo de que no se genere un ciclo infinito (NCSS, 2018).

2.3. Árbol de Decisión

Entre los métodos no paramétricos de predicción tanto para regresión como clasificación están los árboles de decisión, este método también es conocido como Árbol de Predicción. Estos son unos de los algoritmos más utilizados debido a su sencillez y a su fácil implementación. Tipicamente pueden ser muy sensibles a cambios en los datos (carecen de robustez). Este método trabaja por medio de selección recursiva, considerando la mayor parte de la información proporcionada por las variables como herramienta para clasificar la muestra (Hernández, 2004; Orea et al., 2005). Dentro de los métodos de selección por medio de árbol de decisión, existen bastantes metodologías de uso, entre ellas destacan Cart, Chaid, Chaid exhaustivo, Quest, C4.5. Para comenzar con la clasificación de las variables, se generarán particiones recursivas hasta llegar a la estructura final del árbol. En un principio se selecciona la raiz del árbol, en este paso el algoritmo examina una por una cada variable explicativa, y selecciona la variable que mayor significancia tiene para el modelo, siguiendo con sus hojas que serán determinadas de la misma forma, dependiendo del aporte de cada una de las variables en el modelo. Para poder realizar la partición es necesario determinar un punto de corte que será denotado por t, este punto divide a la muestra de tal manera que la variabilidad en el conjunto sea mínima (Hernández, 2004).

Árbol de Decisión Binario

Para este proyecto se utilizará el método Cart de selección binaria. Este método comienza dividiendo la muestra en dos segmentos o subconjuntos, en función de la variable más significativa para el conjunto de predictoras. Cada uno de estos subgrupos es homogéneo dentro y heterogéneo entre ellos. Es decir para la división de la muestra es necesario definir un criterio para la mejor división de esta variable, que será separada por el punto de corte t . Este método fue introducido por Breiman, Friedman, Olshen y Stone, en el año 1984. En esta investigación se consideran variables tanto de tipo cuantitativo como cualitativo. Para esta investigación, la variable respuesta considerada es una variable continua. El modelo de árbol de regresión planteado en este proyecto es de tipo binario, un árbol de tipo binario es dividido en dos nodos hijos o un nodo terminal (Friedman, 1979; Lebart et al., 1995).

Construcción de un Árbol de Decisión para una Variable Continua

Lebart et al. (1995), cuando se refiere al árbol de decisión, lo plantea en términos de un conjunto de variables predictivas $X_1, X_2, ...X_p$, asociadas a una variable dependiente Y.

- Sean $X_1, X_2, ...X_p$ el conjunto de predictoras.
- En primer lugar, el método examina cada una de las variables explicativas, seleccionando la que mayor información entrega (la más significativa para el modelo). Esta variable, X_j, será llamada raíz o nodo padre.
- Con esta variableX_j ya seleccionada, se realizan todas las posibles divisiones, de tal manera que $X_j < t$, donde t representa el corte del nodo padre y corresponde a cualquier valor de la variable seleccionada X_j. Esta división separa la muestra en dos segmentos: $R_l(t)$ es el subconjunto que contiene a todos los valores que satisfacen $X_j < t$ y $R_r(t)$ corresponde a todos los valores restantes de X_j que se encuentran a la derecha de t. Este procedimiento de división de la variable en dos subconjuntos es realizado m veces. Así se obtienen m divisiones, seleccionando la mejor división por medio de la suma de cuadrados residual

(RSS). Este procedimiento, que será detallado más adelante, es realizado para cada una de las j variables predictoras en cada uno de los suconjuntos R_t obtenidos, con el objetivo de mejorar el modelo.

- Este procedimiento recursivo se detiene cuando ya no se pueden realizar más divisiones, ya sea porque la cantidad de observaciones en el nodo es muy pequeña o porque ya no se necesitan más divisiones.

Criterio de mínima varianza o Reducción de la varianza

Para seleccionar el subconjunto que produce la menor varianza, es necesario analizar cada una de las m divisiones. Este método, por lo general es utilizado cuando la variable respuesta es continua (árbol de regresión continuo). "La reducción de la varianza de un nodo "t" se define como la reducción total de la varianza de la variable de destino X_j debido a la partición en este nodo"(Breiman, 2017). La obtención de la varianza para la división de la variable por el nodo "t", se define de la siguiente forma:

$$v(m,t) = \left(\frac{n_l}{n}S_l^2 + \frac{n_r}{n}S_r^2\right),$$

donde n corresponde al tamaño de muestra de la variable X_j analizada, n_l es el tamaño de muestra de la parte derecha de la m−ésima subdivisión realizada por el noto "t" y n_r corresponde al tamaño de muestra de la parte izquierda de esta división. Luego el nodo es seleccionado, en función de la mínima varianza, es decir, se selecciona el nodo que genere la mímima variabilidad para la división de X_j.

$$v(t) = min\{v(m,t)\},$$

Este procedimiento es realizado para cada X_j seleccionada del conjunto de variables predictoras.

2.4. Regresión Bridge

Regresión Bridge es un método de penalización propuesta por Frank Friedman en 1993 (Frank et al., 1993; Fu, 1998). Este método es útil en cualquier situación donde existan problemas de multicolinealidad o sea necesario

seleccionar variables (Orea et al., 2011).

Los estimadores son obtenidos minimizando la suma de cuadrados ordinarios sujeto a la siguiente restricción:

$$\sum_{j=1}^{p} |\beta_j|^{\gamma} \leq t, \quad \gamma > 0,$$

donde t corresponde al nivel de restricción impuesta a los parámetros y se asume un valor pequeño. La regresión Bridge, tiene por objetivo elegir un estimador $\hat{\boldsymbol{\beta}}$ de β penalizado, tal que al minimizar la suma de cuadrados ordinarios, las desviaciones de la variable respuesta sean reducidas. Por lo tanto, la estimación de los coeficientes del vector $\boldsymbol{\beta}$, surge de:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{X}_i^{'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma} \right\}, \tag{2.1}$$

donde y_i representa cada observación, perteneciente al vector de variable respuesta $\mathbf{Y}$; $\mathbf{X}_i^{'}$ es el vector fila asociado a cada observación; $\boldsymbol{\beta}$ corresponde al vector de parámetros y $\lambda \sum_{j=1}^{p} |\beta_j|^{\gamma}$ corresponde a la regularización, que puede ser expresado como, $pen_{\gamma}(\lambda, \beta)$.

El método de regresión Bridge, selecciona variables cuando $0 < \gamma \leq 1$ y reduce los coeficientes cuando $\gamma > 1$. Cabe destacar que, este método no entrega soluciones para cualquier valor de γ.

Regresión Bridge tiene diversos casos particulares, dependiendo del valor que se asocie a γ (Orea et al., 2011). Estos casos serán detallados a continuación.

2.4.1. Regresión Ridge

Esta regresión es un caso especial de Bridge y ocurre cuando el parámetro regularizador $\gamma = 2$. Este tipo de regresión se destaca por eliminar variables que se encuentran correlacionadas, procedimiento que realiza a través de una penalización, la cual afecta a este tipo de variables, contrayéndolas a cero. Este estimador fue propuesto para dar solución a problemas de multicolinealidad (Bühlmann et al., 2011). Así, el estimador de Ridge tiene la misma

forma del estimador de Bridge a diferencia del parámetro de regularización. Es así como la penalización para este caso es, $pen_2(\lambda, \beta)$.

2.4.2. Regresión LASSO

En el año 1996, Robert Tibshirani introdujo LASSO, acrónimo de Least Absolute Srhinkage and Selection Operator (operador de selección y contracción mínima absoluta). Este tipo de regresión al igual que la regresión Ridge, es un caso especial de la regresión Bridge, en el que el parámetro regularizador $\gamma = 1$, que corresponde a la restricción de la norma l_1, $||\beta_j||_1 = \sum |\beta_j|$, (Orea et al., 2011). Este método tiene por objetivo seleccionar variables y, al igual que el caso anterior, realiza una penalización a cada una de las variables contrayendo algunas a cero, por medio de un parámetro regularizador que se denota por λ. Este parámetro se obtiene a través de validación cruzada (Bühlmann et al., 2011). Este método se verá en detalle en los próximos capítulos.

El estimador asociado a LASSO, se define tal como el caso general de Bridge, donde $\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^{'}\boldsymbol{\beta})^2$ corresponde al estimador de mínimos cuadrados, pero esta vez sujeto a la regularización l_1 de LASSO, $pen_1(\lambda, \beta)$.

2.4.3. Regresión Elastic-Net

El método de regresión Elastic-Net (E-N) corresponde a una generalización de regresión LASSO y Ridge, donde ocupa una combinación de las penalizaciones l_1 tipo LASSO y l_2 de tipo Ridge (Bühlmann et al., 2011). De esta forma, el estimador E-N es realizado minimizando la suma de cuadrados ordinarios, pero esta vez sujeto a ambas regularizaciones en conjunto, por lo tanto, $pen_{\alpha,\lambda}(|\beta_j|) = \lambda\{(1-\alpha)||\beta||_1 + \alpha||\beta||^2\}$.

Los modelos de regresión penalizada entregan métodos estadísticamente atractivos, para la predicción de modelos, provenientes de grandes cantidades de variables. A estas variables se les analizan sus características buscando ajustar un buen modelo (Sill et al., 2014).

2.4.4. Métodos de Regularización

Como se puede observar en estos métodos de regularización, se añade una penalización al vector de parámetros y dependiendo del tipo de regresión que se utilice, sólo encogerá el vector de parámetros o incluso seleccionará variables llevando algunos de estos coeficientes a cero.

Capítulo 3

Selección del Mejor Modelo

3.1. Validación Cruzada

La selección del mejor modelo, depende del valor que se le asigne al parámetro $\lambda \geq 0$. Este parámetro por lo general es seleccionado por medio de validación cruzada o bootstrap, pero como se mencionó anteriormente para obtener el mejor valor de λ en este trabajo, se utilizará validación cruzada. Esta técnica es usada para estimar el nivel de ajuste que tiene un modelo.

La realización de esta herramienta consiste en dividir en dos partes el conjunto de datos, los cuales se denominan conjunto de prueba y conjunto de entrenamiento; estos conjuntos son independientes entre sí.

En este trabajo, en las siguientes secciones, se presentarán dos tipos de validación cruzada.

3.1.1. Validación Cruzada Dejando una Observación Afuera

El primer método se denomina validación cruzada dejando una observación fuera y es el método más utilizado. Consiste en extraer una observación de la muestra como dato de prueba, y todas las observaciones restantes corresponden a los datos de entrenamiento (Arlot et al., 2010). Suponiendo que

se cuenta con n observaciones se modela en base a las $n-1$ observaciones restantes. Es así como se obtiene el error de predicción del modelo para la observación de prueba y el valor estimado mediante las $n-1$ observaciones verificadas en el modelo. La diferencia entre estos valores es elevado al cuadrado. Este procedimiento se realiza para cada una de las n observaciones en la muestra y la suma de estos errores entrega como resultado el error cuadrático medio de las n observaciones.

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

donde y_i representa a cada observación y $\hat{y}_i$, representa a cada valor estimado con los $n-1$ restantes.

Cada una de las n obsevaciones es analizado por este método. Luego son comparados y se selecciona el que tiene menor Error Cuadrático Medio (puede trabajarse en base a otra medida de error). El problema que presenta este tipo de validación cruzada es que al tener que evaluar el modelo n veces requiere de un mayor tiempo de análisis, lo que produce un gasto computacional demasiado grande (Arahal et al., 2006).

3.1.2. Validación Cruzada con k Particiones

Para este tipo de validación, el conjunto de datos se divide en k subconjuntos disjuntos. Uno de estos subconjuntos se deja como conjunto de prueba y los $k-1$ subconjuntos restantes como datos de entrenamiento (Bengio et al., 2004). Este procedimiento se realiza para cada uno de los subconjuntos, utilizando en cada iteración un subconjunto como conjunto de prueba. Debido a esto, el procedimiento se realizará k veces. En cada una de las k iteraciones se obtendrá un error de predicción que está asociado a cada conjunto de prueba.

Realizar este método tiene como inconveniente el gasto computacional necesario para verificar el modelo. Este gasto es proporcional al número de particiones consideradas.

En cada iteración se toma el conjunto de entrenamiento que corresponde a los $k-1$ subconjuntos y se verifica el modelo, comparándolo con los

datos de prueba. De aquí se obtiene el error de estimación para cada una de las iteraciones. Aquí, el ECM se define como:

$$ECM = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2,$$

donde y_i corresponde a cada una de los grupos observados, $\hat{y}_i$ es la estimación realizada con los $k-1$ datos de prueba y k es la cantidad de particiones o grupos realizados.

Validación cruzada permite seleccionar modelos y debido a esto también puede ser utilizado como criterio de selección de variables, familias de aproximaciones, tamaño de muestra, etc. (Arahal et al., 2006).

Se debe tener en cuenta que además de estos dos métodos existe la validación cruzada de tipo aleatoria, la cual no será considerada en el presente proyecto.

Capítulo 4

Métodos de Regularización

Durante los últimos años se han dado a conocer distintas técnicas de selección de variables y estimación de los efectos que estas tienen en una determinada respuesta. Debido a la eficiencia que presentan estos métodos, tanto en selección de variables cuanto en regularización de los coeficientes, es que están siendo utilizados para selección de modelos y estimación de los parámetros asociados a estos modelos.

En general este tipo de métodos son muy útiles, cuando hay problemas de dimensión de variables.

Fan et al. (2001) propusieron algunas propiedades que debiesen cumplir los métodos de penalización:

1. Consistencia del estimador de parámetros $\hat{\boldsymbol{\beta}}$: se dice que el estimador es consistente si el valor estimado de éste, converge a su valor verdadero cuando el tamaño de la muestra tiende a infinito, es decir,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} \xrightarrow{p} \beta, \qquad n \longrightarrow \infty \tag{4.1}$$

2. Normalidad Asintótica del vector de parámetros estimado:

$$\sqrt{n}(\hat{\boldsymbol{\beta}_{\mathbf{1}\boldsymbol{n}}} - \boldsymbol{\beta}_{\mathbf{1}}) \longrightarrow N(0, \mathbf{V}), \tag{4.2}$$

   donde $\mathbf{V}$ corresponde a una matriz de dimensión n×p, con n igual al número de parámetros seleccionados.

3. consistencia del modelo: es decir que la probabilidad de que el modelo seelccionado sea el óptimo es 1.

$$P(\hat{M} = M)) \xrightarrow{p} 1, \tag{4.3}$$

donde $\hat{M}$ representa al modelo estimado.

4. Esparsidad: la selección de variables debe ser automática en caso de que los coeficientes asociados a la variable sean pequeños, es decir, la probabilidad de que un coeficiente no seleccionado sea 0, es 1.

$$P(\hat{\boldsymbol{\beta}} = 0_n)) \xrightarrow{p} 1, \tag{4.4}$$

5. Insesgamiento: debe presentar un sesgo bajo, sobretodo cuando los coeficientes presentan valores grandes.

En este capítulo se darán a conocer dos métodos de selección de variables.

4.1. Regresión LASSO

En el año 1996, Tibshirani introdujo LASSO. Como se mencionó en el Capítulo 2, este método es un caso especial de regresión Bridge ($\gamma = 1$), por lo tanto, al igual que el método introducido por Friedman, minimiza la suma de cuadrados ordinarios, sujeto a la siguiente penalización:

$$\sum_{j=1}^{p} |\beta_j| \leq t.$$

Para poder estimar el vector de parámetros, se contrae el estimador de mínimos cuadrados ordinarios $\hat{\beta}_{ols}$, combinado con la restricción de LASSO, a cero y potencialmente el conjunto de coeficientes $\hat{\beta}_j = 0$, para algún j, con $j = 1, \ldots, p$. Además, para la realización de este proceso, se considera un comportamiento asintótico, de los estimadores de regresión (knight et al., 2000). Es decir, los valores de coeficientes pequeños serán contraídos a cero de manera inmediata, cumpliendo con el supuesto de esparsidad.

Este método no sólo es útil por tener mayor precisión en problemas de predicción, sino que al ser un proceso discreto, también realiza una selección de variables de forma continua, debido a la contracción que realiza para algunos de los coeficientes de regresión a cero. Esto genera modelos con menor cantidad de variables, que a la vez tienen menor dificultad de interpretación.

LASSO es menos variable que otros métodos de selección, tales como, los métodos de selección por pasos stepwise, backward o forward. Todos estos métodos no son óptimos cuando $p >> n$. Además son considerados inestables debido a que cuando se producen pequeños cambios en los datos, el resultado de la selección de variables varía demasiado.

Los métodos de regularización puede ser utilizados en conjunto con otras herramientas de construcción de modelos. Por otro lado requieren que las variables regresoras se encuentren estandarizadas, para evitar que la penalización varíe por cambios de escala (Tibshirani, 1997).

Como LASSO es un método general, puede ser utilizado en una gran cantidad de modelos. El caso más simple es un modelo lineal ortonormal, en el que el estimador de mínimos cuadrados ordinarios corresponde al estimador de máxima verosimilitud. En este modelo, se considera un número de variables igual al número de observaciones. Este es una caso particular del estimador LASSO, el cual es llamado "soft thresholding"(estimador suave) (Fu, 1998). En este caso LASSO coincide con "soft thresholding", cuyo estimador tiene la siguiente forma:

$$\hat{\theta}_S = signo(\bar{y})(|\bar{y}| - \eta_n)_+,$$

donde η_n corresponde al umbral utilizado, signo($\bar{y}$) es -1 cuando $\bar{y}$ es menor que 0, es 0 cuando $\bar{y}$ es cero y es 1 cuando $\bar{y}$ es mayor 1.

En el caso particular donde el número de observaciones coincide con el número de variables el estimador de LASSO tiene la siguiente forma:

$$\sum_{i=1}^{n}(y_i - \theta)^2 + 2n\eta_n|\theta|.$$

Notar que el parámetro de regularización λ de LASSO corresponde a $2n\eta_n$ (Tibshirani, et al., 1997).

Supongamos un conjunto de variables pertenecientes a una regresión lineal simple, donde la matriz de covariables y el vector de variable respuesta son denotados por:

$$\mathbf{X} = [x_1, ...x_p] \qquad \text{e} \qquad \mathbf{Y} = [y_1..., y_n]^{'},$$

donde $x_j = (x_{1j}, \ldots, x_{nj})^T, \quad j = 1, \ldots, p.$

Además se considera un modelo de regresión lineal múltiple en forma matricial, como fue visto en el Capítulo 2:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n), \tag{4.5}$$

donde,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} y \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

La variable respuesta corresponde al vector $\mathbf{Y}_{n\times 1}$, la matriz de variables se define como $\mathbf{X}_{n\times p}$, el vector de parámetros se denota por $\boldsymbol{\beta}_{p\times 1}$ y el error de estimación por $\boldsymbol{\epsilon}_{n\times 1}$. Además, que las p covariables $x_i \in \mathbf{X}$ son variables aleatorias estandarizadas, por lo tanto, tienen media cero y la varianza es constante e igual a 1, esto es:

$$\frac{\sum_{i=1}^n x_{ij}}{n} = 0, \qquad \frac{\sum_{i=1}^n x_{ij}^2}{n} = 1, \quad para \quad j = 1, ..., p.$$

Asumiendo que se cumplen todos los supuestos antes mencionados, se realiza la construcción del algoritmo LASSO, la cual se define como (James

et al., 2013; Tibshirani, 1996):

$$f(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{X}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

En caso de que el número de variables sea igual al número de observaciones, el intercepto será cero. Se observa que regresión LASSO, corresponde al estimador de mínimos cuadrados ajustado a una penalización, denotada por:

$$\lambda \sum_{j=1}^{p} |\beta_j|,$$

donde λ es un parámetro de penalización.

El tamaño del parámetro λ controla la cantidad contraída para cada estimador $\hat{\boldsymbol{\beta}}$, pudiendo generar soluciones más estables. La elección de este parámetro λ es un problema importante. La elección de este debe ser a partir de los datos, por ejemplo, a través de validación cruzada o bootstrap como ya fue mencionado.

El algoritmo de penalización LASSO es un método muy efectivo para seleccionar modelos continuos, debido a que toma las mejores características de la selección de variables. Es por esto que autores tales como Claerbout et al. (1973), Taylor (1979), Santosa et al. (1986), Tibshirani (1996), Fu (1998), Chen (1998), Daubechies (2004), Wu (2008), han utilizado penalización LASSO en ambos ajustes l_1 y l_2.

4.1.1. Descripción de λ y t

Como ya se mencionó en el apartado anterior, el estimador de LASSO tiene la propiedad de seleccionar las variables en el sentido que $\hat{\beta}_j(\lambda) = 0$, dependiendo del valor que toma λ. Así puede pensarse como un $EMC(\hat{\boldsymbol{\beta}})$.

Para valores grandes de λ los $\beta_j's$ se contraen a cero. Es por esto que, dependiendo del valor que asuma λ, será la intensidad de la penalización que

se aplica a cada coeficiente β_j, entregando soluciones esparsas (nulas para algunas y no nulas para las restantes).

El parámetro λ controla la penalización en el proceso. A mayor λ, mayor es la penalización en los coeficientes de regresión. En el caso de que λ sea cero, coincide con el estimador de mínimos cuadrados. Si $\lambda \to \infty$, se tiene que $\hat{\beta}(\lambda) \to 0$.

Es por esto que la cantidad de variables seleccionadas depende netamente del valor que se le asigna al parámetro de penalización λ. En caso de reducir λ, automáticamente se reduce la penalización, por lo que una mayor cantidad de variables pueden ingresar al modelo (Wu, 2009).

4.1.2. Desventajas del Algoritmo

Los principales inconvenientes que presenta LASSO son:

1. No es posible seleccionar mayor cantidad de variables explicativas que el número de observaciones.
2. Cuando hay problemas de multicolinealidad LASSO selecciona sólo una variable entre muchas variables que se encuentran correlacionadas.
3. LASSO no es adecuado cuando no se cumplen las siguientes propiedades: consistencia del estimador de parámetros, normalidad asintótica, consistencia del modelo y esparcidad.

Para solucionar las dos primeras desventajas se puede utilizar E-N, debido a que además de seleccionar variables se preocupa de la multicolinealidad que pudiese estar presente en las covariables (Zou, 2005).

Para dar solución a la última desventaja, en que LASSO no puede entregar modelos eficientes, se propone LASSO Adaptative. Este procedimiento es muy útil para corregir los problemas de sobreestimación que presenta LASSO, pues éste método tiene las propiedades de oráculo mencionadas en el comienzo de este capítulo (Caner, 2010).

4.1.3. Estimación de la Varianza

El estimador de regresión LASSO no entrega directamente un estimador para la varianza del error σ^2.

Cuando el número de variables excede el número de observaciones, la estimación de la varianza no se puede realizar por medio de mínimos cuadrados, debido a que los errores tienden a ser grandes y es por esto que el estimador es inestable. Es decir no existe el EMCO, pues $\nexists\,(X'X)^{-1}$. Debido a esto, muchos autores han propuesto distintas formas de estimar la varianza del error σ^2, y así poder observar la significancia que el vector de parámetros presenta.

Se puede construir un estimador usando la suma de cuadrados residuales y los grados de libertad asociados a LASSO. Un estimador que parece ser útil es (Reid, 2013):

$$\sigma^2 = \frac{1}{n - \hat{s}_{\hat{\lambda}}} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' \boldsymbol{\beta}_{\hat{\boldsymbol{\lambda}}})^2,$$

donde λ como ya fue mencionado es el parámetro de penalización y es seleccionado por medio de validación cruzada y $\hat{s}_{\hat{\lambda}}$ corresponde al número de variables seleccionadas por LASSO (número de elementos distintos de cero en el vector de parámetros $\boldsymbol{\beta}_{\hat{\boldsymbol{\lambda}}}$).

El problema, es que estas estimaciones de la varianza del error propuestas por diversos autores, pueden provocar un sesgo muy grande cuando la muestra es muy pequeña. Además como el estimador de σ^2 depende del valor λ ya seleccionado por validación cruzada, a medida que este aumenta, mayor es el sesgo y menor es la varianza (Reid, 2013).

Este no es el único método de estimación de la varianza del error. Como alternativa se han desarrollado otros estimadores de la varianza, usando una reparametrización propuesta por Bühlmann et al. (2011).

4.2. Regresión E-N

Este algoritmo fue introducido por Zou (2005), como generalización de regresión LASSO, debido a las desventajas que presente éste.

Es bastante utilizada cuando el número de variables es mayor que el número de observaciones. Al igual que LASSO selecciona variables contrayendo algunos coeficientes a cero.

Presenta las mismas propiedades descritas en la sección anterior, pero además de aplicar una penalización de tipo l_1 $(\lambda|\beta_j|)$, aplica una segunda penalización de tipo l_2 $(\lambda|\beta_j|^2)$, correspondiente a regresión Ridge. Esta penalización es muy útil en caso de existir multicolinealidad (Sill et al., 2014). Minimiza la suma de cuadrados ordinarios, sujeto a dos penalizaciones:

$$p_{\alpha,\lambda}|\beta_j| = \sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2 + \lambda\left\{\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right\},$$

donde λ es el parámetro de penalización obtenido mediante validación cruzada y α con su complemento $(1-\alpha)$ forman una combinación convexa sobre LASSO y Ridge, respectivamente. Los valores para el parámetro regularizador, correspondiente a cada penalización, son los mencionados anteriormente.

Por lo general este método tiende a seleccionar más variables que regresión LASSO (Sill et al., 2014).

4.2.1. Elección de λ y α

El valor del parámetro λ es seleccionado por validación cruzada al igual que regresión LASSO. Dependiendo del valor de este parámetro, es la penalización aplicada al vector de parámetros por medio de regresión E-N, recordemos que el valor de $\lambda \geq 0$.

El valor del parámetro $\alpha \in [0, 1]$. En caso de que el valor que se designa sea pequeño, esto indicará que se acerca a regresión Ridge. En cambio si este valor es cercano a uno, esto indicará que E-N se comporta como regresión LASSO (Sill et al., 2014).

Capítulo 5

Resultados

5.1. Descripción de los Datos

Este trabajo de investigación se enfocó en el problema de la DM en mujeres adultas mayores y las variables relacionadas con esta. La muestra reclutada en el año 2012 para el estudio está constituida por 96 mujeres AM, que realizaban alguna actividad física, en dos casas de adulto mayor pertenecientes a la ciudad de Valdivia. Para cada una de estas mujeres se registraron 93 variables, de carácter cuantitativo y cualitativo, incluyendo variables de tipo demográfico, biomédicas, antropométricas, de clasificación de la población y de DM.

La variable de interés para el estudio es el TM6M, el cual fue realizado siguiendo la recomendación del consenso internacional realizado por la Society on Sarcopenia, Cachexia and Wasting Disorders. En este concenso realizado en Roma el año 2009, se definió que un AM posee Sarcopenia con Movilidad Limitada cuando "camina menos de 400 metros durante la prueba de caminata en 6 minutos TM6M; y presenta una pérdida de masa muscular significativa en comparación a un grupo de personas con edades entre 20 y 30 años". El año 2011 se determinó que una mejora funcional y clínicamente significativa en DM, es apreciable con un cambio de al menos 50 metros durante el TM6M (Morley et al., 2011). Esta prueba se aplicó en un lugar amplio, donde se delimitaron 15 mts lineales. Las adultas mayores debían caminar ida y vuelta durante estos seis minutos de examen, mientras se les incentivó constantemente a continuar la caminata (Figura 5.1). Mientras las participantes caminaban, se les tomaban otras medidas, tales como

Figura 5.1: Corredor plano del test de marcha seis minutos.

Frecuencia Cardiaca (FC), Escala de Esfuerzo de Borg y Presión Arterial (PA) (previo al test, durante el TM6M y post test). La severidad de DM se mide en grados (Morley et al., 2011), de a cuerdo a intervalos de 50 metros (ver Tabla 5.1).

Las demás mediciones se regitraron previo a la aplicación del TM6M. Es así como se definió como activas a las mujeres que realizaba actividad moderada tres veces a la semana o que realizaban actividad intensa al menos una vez a la semana, como lo define la Encuesta Nacional de Salud 2014 (ENS 2014). La DMII se basó en diagnóstico previo, La FPM se utilizó para clasificar a las participantes como Dinapénicas, basándose en un grupo de referencia de estudiantes de la Universidad de Valdivia que arrojó como punto de corte el valor 23,5 Kg (Navia et al., 2012). Las adultas mayores que presentaban un PCI mayor a 80 cm fueron consideradas como Obesas (ENS 2014) y la clasificación Obesa Dinapénica (OD), indica la presencia de ambas enfermedades en conjunto.

Severidad de la disfunción motora	
Grado	Metros recorridos en 6 minutos
DM grado 0	Mayor o igual a 400 metros.
DM grado 1	Entre 350 y 399 metros.
DM grado 2	Entre 300 y 349 metros.
DM grado 3	Entre 250 y 299 metros.
DM grado 4	Menor a 250 metros.

Tabla 5.1: Grado de disfunción motora.

5.2. Análisis de los Datos

Para seleccionar las variables apropiadas para la predicción de DM, se utilizaron cuatro técnicas de selección de variables, las cuales fueron descritas en el capítulo anterior. Dos de estas técnicas constituyen métodos clásicos (Stepwise y árbol de decision) y las otras son métodos de regularización para la selección de variables através de penalización de parámetros (LASSO, recomendado cuando la cantidad de variables es muy alta y Elastic Net, recomendado cuando además, existe multicolinealidad en el conjunto de predictoras)(Bühlmann et al., 2011; Fu, 1998).

5.2.1. Selección por medio de Stepwise

Para la aplicación de Stepwise se utilizó un nivel de significación de 0.15, tanto para el ingreso como para la salida de variables. Del conjunto de 93 variables, fueron seleccionadas 9 variables, las cuales se presentan en la Tabla 5.2.

Stepwise			
Variable	Sensibilidad	Especificidad	Correctas
FPM	97	94	85
Pci (cm)			
FC4'			
Diabetes Mellitus			
Dinapénicas			
AMM3			
MM7			
Sistólica pre			
Artrosis			

Tabla 5.2: Características de las variables eleccionadas mediante Stepwise.

5.2.2. Selección por Medio de Árbol de Decisión.

Para la aplicación de Árbol de Decisión, se utilizó el criterio de mínima varianza que trabaja por medio de mínimos cuadrados. Además se especificó un tamaño mínimo de muestra 10, para la división del nodo. Del conjunto de 93 variables, fueron seleccionadas 9 variables, las cuales se peresentan en la Tabla 5.3.

Árbol de Decisión			
Variable	Sensibilidad	Especificidad	Correctas
FPM	59	89	74
Diabetes Mellitus			
Edad			
Peso			
Altura (cm)			
FCM			
Activa			
AMM3			
HU9			

Tabla 5.3: Características de las variables seleccionadas mediante Árbol de Decisión.

5.2.3. Selección por Medio de Regresión Elastic Net (E-N).

Para la aplicación de regresión E-N, se utilizó validación cruzada, dejando una observación afuera. Del conjunto de 93 variables, fueron seleccionadas 8 variables, las cuales se peresentan en la Tabla 5.4

Elastic Net			
Variable	Sensibilidad	Especificidad	Correctas
FPM	97	98	85
Pci (cm)			
Diabetes Mellitus			
Ob.Dinap 2014			
FC4'			
Edad			
Dinapénicas			
AMM5			

Tabla 5.4: Características de las variables seleccionadas mediante E-N.

5.2.4. Selección por Medio de Regresión LASSO.

Para la aplicación de regresión LASSO, al igual que el método anterior, se utilizó validación cruzada dejando una observación afuera. Del conjunto de 93 variables, fueron seleccionadas 5 variables. Luego se aplicó regresión convencional, donde se verificó que todas las variables eran significativas, por lo cual, el número de variables seleccionado finalmente fueron las 5 seleccionadas por LASSO, las cuales se presentan en la Tabla 5.5.

LASSO			
Variable	Sensibilidad	Especificidad	Correctas
FPM	100	98	86
Pci (cm)			
Diabetes Mellitus			
Ob.Dinap 2014			
FC4'			

Tabla 5.5: Características de las variables seleccionadas mediante LASSO.

Desde la Figura 5.2 se puede observar la curva de validación cruzada para el método de LASSO (a la izquierda) y para E-N (a la derecha). En ambas imágenes podemos ver la curva de validación cruzada representada por los puntos de color rojo, junto a sus respectivas desviaciones estándar.

Para el caso de E-N, el valor de lambda.min (linea punteada que se encuentra a la izquierda de cada figura y que representa la mínima desviación) presenta mayor variabilidad que en el caso de la curva obtenida por medio de regresión LASSO. Además, se puede observar que, dependiendo del valor asociado a este lambda.min, son seleccionadas las variables que se encuentran representadas en la parte superior del gráfico por los grados de libertad. El número de variables seleccionadas no coincide para ambos métodos, seleccionándose 5 variables por medio de LASSO y 8 variables a través de regresión E-N.

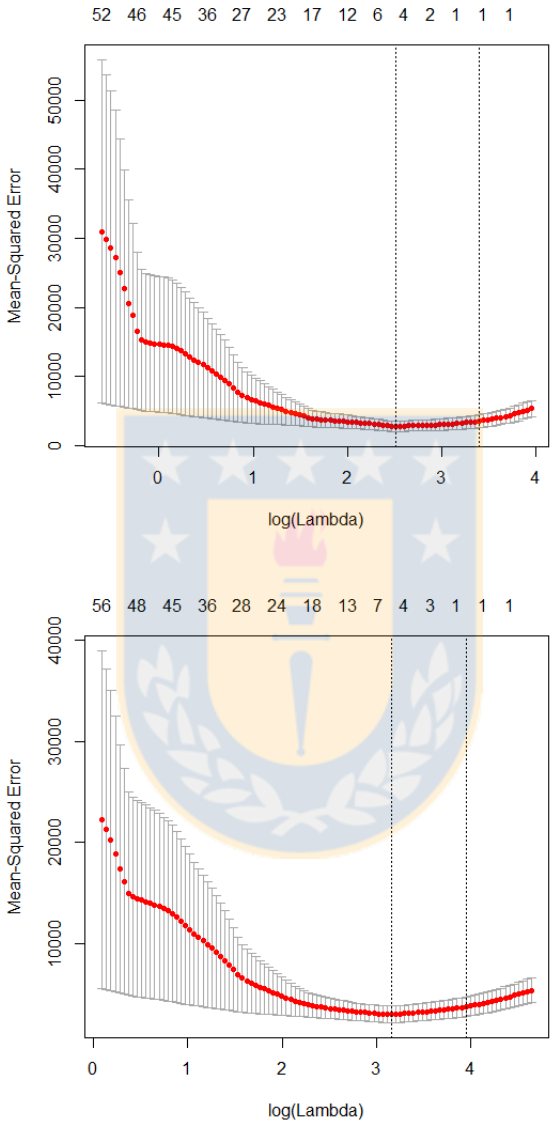

Figura 5.2: Curva de validación cruzada por medio de regresión LASSO y E-N aplicadas al conjunto de datos de DM.

El valor del parámetro de regularización λ seleccionado fue el entregado por medio de validación cruzada (lambda.min), con un valor de 11.79362 y

22.51517, respectivamente para cada tipo de regularización. Cabe destacar que, como fue mencionado anteriormente, la metodología utilizada en E-N no sólo se preocupa de seleccionar variables encogiendo los coeficientes a cero, sino que además es útil cuando existen problemas de multicolinealidad. Ambos modelos seleccionaron practicamente las mismas variables. Como se mencionó anteriormente el método de penalización LASSO, cuando existen predictoras muy correlacionadas selecciona cualquiera de ellas y no siempre la más adecuada; E-N no presenta este problema, por lo tanto la similitud entre los dos conjuntos de variables seleccionados nos brinda confianza de los resultados proporcionados por LASSO.

Desde la Figura 5.3, podemos ver el comportamiento de las variables tanto para regresión LASSO como E-N. Las curvas representan la trayectoria que sigue cada variable perteneciente al conjunto de datos.

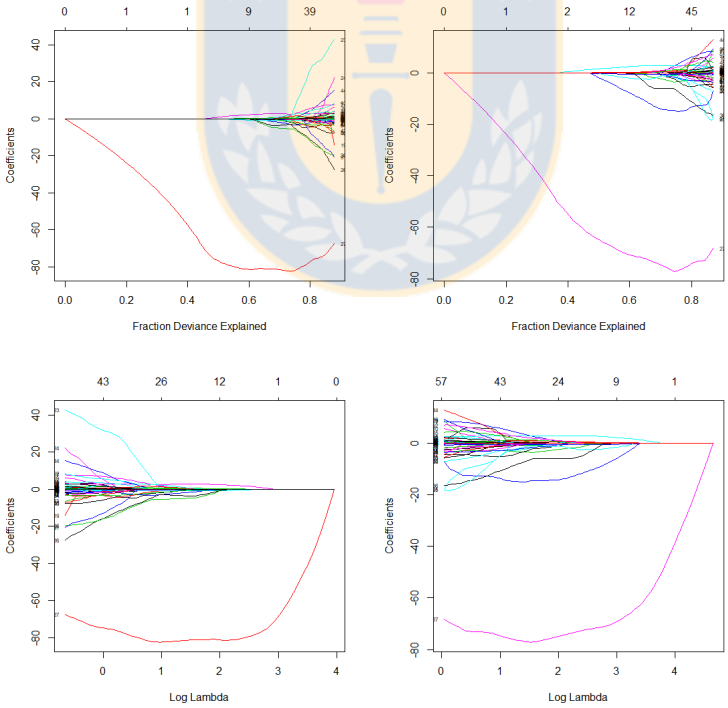

Figura 5.3: Comportamiento de los coeficientes, por medio de regularización LASSO y E-N, aplicados a la muestra de DM.

Las dos figuras superiores (ver Figura 5.3), muestran el porcentaje de devianza explicada por las variables seleccionadas. Como se puede observar, en el caso de LASSO se seleccionaron 5 variables y E-N seleccionó tres más, siendo el porcentaje de desvianza explicada aproximadamente un 60 % para ambos casos. Ambos métodos seleccionaron una cantidad similar de variables. Para conocer cual de estos modelos es el que presenta un mejor ajuste, se realizó una comparación de ambos por medio del criterio de Akaike corregido (ver Tabla 5.6), en cada modelo se puede observar el peso en la ponderación final. Este peso puede ser interpretado como la probabilidad de que el modelo evaluado sea mejor en términos de ajuste y por parsimonia. El modelo asociado al algoritmo de LASSO tiene un valor de AICc menor que el implementado por E-N, la probabilidad asociada para ser un buen modelo es de 0.61 y además LASSO seleccionó menor cantidad de variables.

Criterios			
Método	K	AICc	AICcWt
LASSO	5	832.57	0.61
E-N	8	835.43	0.14
Propuesto	4	839.88	0.02
LASSO sin FCM4	4	834.47	0.23

Tabla 5.6: Criterios de selección.

5.2.5. Resumen de Resultados

Los cuatro métodos predictivos de DM coinciden en seleccionar las variables DMII y FPM. Stepwise incluye además Frecuencia Cardíaca en el minuto 4 (FCM4), PCI, Clasificación de Dinapenia, Presión Arterial y Artrosis. El árbol de predicción incluye, además, las variables Edad, Peso, Talla, Frecuencia Cardíaca Media y Clasificación de Activas.

Resumen de Métodos				
Método	N°	Sensibi.	Especifi.	Correctas
Stepwise	9	97	94	85
Árbol	9	59	89	74
E-N	8	97	95	81
LASSO	5	100	95	84
Propuesto	4	0	100	67
Sin FCM4	4	97	97	81

Tabla 5.7: Selección del mejor método de selección de variables.

Por otra parte, y en virtud de las dificultades para la aplicación clínica de los modelos se seleccionaron cuatro variables, considerando la rapidez y factibilidad de obtención de sus mediciones, las cuales constituyeron un modelo alternativo que incluía: FPM, IMC, presencia de DMII y Obesidad. Se evaluó el poder predictivo de estas, obteniendo un bajo rendimiento del modelo, con una sensibilidad de 0 % (Tabla 5.7), siendo inviable proponer un modelo predictivo de DM con las variables seleccionadas.

El modelo que presenta una mejor predicción de la enfermedad y los niveles de ésta es LASSO, seleccionando las variables, PCI en cm, FPM, clasificación de OD, DM y FC4M. Con estas variables se realizó un análisis de regresión lineal, para el cuál todas las variables resultaron ser significativas. En conclusión, los factores de riesgo que se utilizaron para predecir la DM son PCI en cm, FPM, clasificación obesas dinapénicas, diabetes Mellitus y FC4M.
Entonces el modelo inicial para predecir DM incluye las siguientes variables: PCI(p-valor = 0.00924), FPM(p-valor = 0.00109), Obesidad Dinapenia(p-valor = 0.01910), DMII(p-valor = 4.06e-10) y FCM4(p-valor = 0.04670).
El análisis de poder predictivo de todos los modelos propuestos por cada uno de los cuatro métodos de selección de variables y de los dos propuestos por el investigador, son mencionados anteriormente en la Tabla 5.7. LASSO fue el método con mayor eficiencia predictiva, por ello, con la finalidad de llegar a un modelo de mayor utilidad práctica, se consideró disminuir su eficiencia. En aras de la aplicabilidad práctica del mismo. Para ello, como la variable FCM4 resultaba ser la menos significativa dentro del modelo, se propuso un modelo que no la incluye, y se evaluó el poder predictivo del modelo LASSO, pero esta vez sin FCM4, obteniendo una sensibilidad de un 97 % (es decir la probabilidad de seleccionar correctamente a una AM que tiene DM es de 0.97) y una especificidad de 100 % (es decir, la probabilidad de identificar

a una AM que no presenta DM es 1). Además, se obtuvo un porcentaje de correcta selección de un 84 %(es decir, en el 84 % de los casos las pacientes son clasificadas no sólo en términos de la presencia o ausencia de DM, si no también en términos del grado de dicha disfunción). Es por esto que las variables seleccionadas para realizar la predicción son: PCI, FPM, Obesidad Dinapenia y DMII.

Correlación de TM6M según Obesidad Dinapenia y DMII

Factor	PCI	FPM	FCM4
TM6M	r=-0.33 (p= 0.3262)	r=**0.33** (p= **0.0030**)	r=**0.31** (p= **0.0054**)
No O ni D	r=-0.14 (p=0.8159)	r=0.36 (p=0.4299)	r=**0.90** (p=**0.0387**)
O	r=-0.22 (p=0.1399)	r=0.23 (p=0.1199)	r=**0.30** (p=**0.044**)
NO	r=0.20 (p=0.6038)	r=0.67 (p=0.050)	r=0.49 (p=0.2201)
D	r=**-0.41** (p=**0.0032**)	r=**0.30** (p=**0.0333**)	r=**0.31** (p=**0.0330**)
ND	r=-0.22 (p=0.1399)	r=0.23 (p=0.1199)	r=**0.30** (p=**0.0443**)
OD	r=**-0.50** (p=**0.0004**)	r=**0.30** (p=**0.0448**)	r=**0.32** (p=**0.0354**)
NOD	r=-0.11 (p=0.4283)	r=0.28 (p=0.0483)	r=**0.29** (p=**0.0424**)
DMII	r=**-0.52** (p=**0.0023**)	r=-0.06 (p=0.7242)	r=0.22 (p=0.2457)
No DMII	r=-0.03 (p=0.8370)	r=**0.31** (p=**0.0116**)	r=0.23 (p=0.0653)

Tabla 5.8: Correlación de las variables con TM6M. Correlaciones según obesidad y/o dinapenia y presencia de DMII.

Desde la Tabla 5.8, se observa que para algunos grupos particulares existe asociación significativa entre TM6M y las variables PCI, FPM y FCM4.
En primer lugar, en el grupo completo, se encuentra que existe correlación significativa y directa entre TM6M y las variables FPM y FCM4, (r=0.33, p=0.0030) y (r=0,31, p=0,0054), respectivamente.
Existe correlación significativa y directa entre TM6M y la variable FPM, en el grupo de No Diabéticas (r=0.31, p=0.0116), en el grupo de Obesas Dinapénicas (r=0.30, p=0.0448), en el grupo que no son Obesas Dinapénicas (r=0.28, p=0.0483) y en el grupo de Dinapénicas (r=0.30, p=0.0333).
Existe correlación significativa y directa entre TM6M y la variable FCM4, en el grupo de Obesas Dinapénicas (r=0.32, p=0.0354), en el grupo que no

son Obesas Dinapénicas (r=0.29, p=0.0424), en el grupo de Dinapénicas (r= 0.31, p= 0.033) y en el grupo de No Dinapénicas (r= 0.30, p= 0.0443). Existe correlación significativa e inversa entre TM6M y la variable PCI, en el grupo de Diabéticas (r= -0.5202, p=0.0023), en el grupo de Obesas Dinapénicas (r= -0.50, p=0.0004) y en el grupo Dinapénicas (r= -0.41, p=0.0032) (ver Figura 5.4).

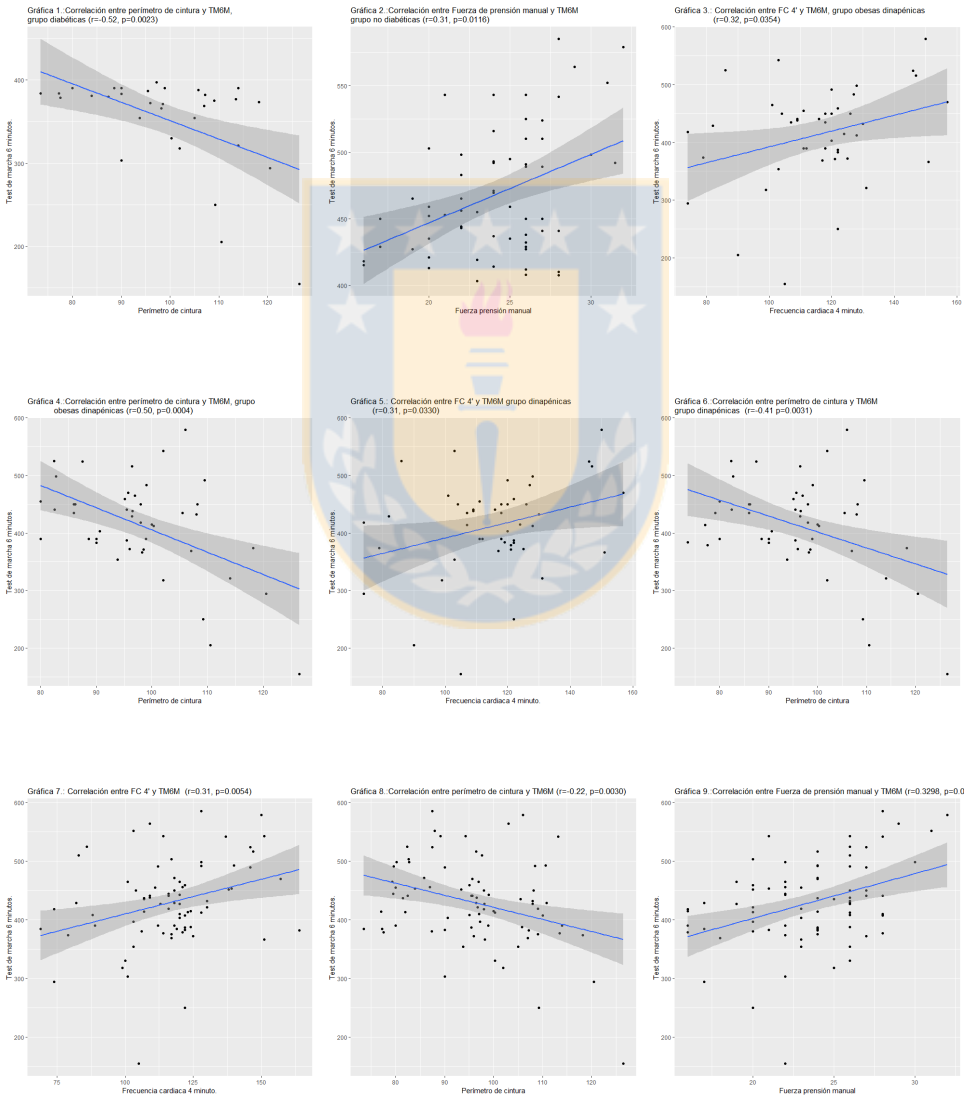

Figura 5.4: Correlación de las variables con TM6M, por grupo.

Además, se obtuvo la razón de chance de padecer DM, así se pudo estimar que una persona que es Obesa Dinapénica tiene 2.25 veces más chance de tener DM que aquellas que no son Obesas Dinapénica, y que una AM que tienen DMII tienen 8385 veces más de chance de presentar DM, que las que no tienen DMII (en la muestra estudiada el 100 % de las personas que padecen DMII, presentan DM,y de las personas que no presentan DMII, ninguna presenta DM). Éste es un factor determinante (ver Tabla 5.9).

TM6M		
Variable	OR	P-Valor
OD	2.25	0.0355
DMII	8385	3.5412E-06

Tabla 5.9: Análisis de Odds Ratio de variables asociadas a disfunción motora.

La Tabla 5.10 presenta el resumen descriptivo porcentual, según condición de variables biodemográficas; obesidad, dinapenia, sedentarismo, DM y DMII; de las 96 participantes, sólo el 5 % No presenta ni obesidad ni dinapenia, un 4.2 % es sólo dinapénica, aproximadamente el 44 % es obesa, el 45 % es OD, y un 32 % presentan DMII y DM a la vez.

Tabla de Frecuencias de OD, DMII y nivel de actividad física en AM.

Condición	Cantidad	%
Dinapénicas no Obesas (G1)	4	4.2
No Obesas ni Dinapénicas(G2)	5	5.2
Obesas no Dinapénicas (G3)	42	43.8
Obesas Dinapénicas (G4)	45	46.9
Diabetes Mellitus	32	33.3
Sedentaria	40	41.7
DM	32	33.3

Tabla 5.10: Porcentaje y frecuencia del total de AM.

La Tabla 5.11, muestra que en el grupo de mujeres AM que no tienen obesidad ni dinapenia, el 60 % es activa y un 20 % tiene DMII. La edad promedio de este grupo fue 68.6 $\pm$ 1.8 años, presentan un perímetro de cintura promedio de 77.8 $\pm$ 2.6 cm, la FPM promedio fue de 23.8 $\pm$ 1.8 kg, la FCM4

promedio fue 111 ± 25.5 y el tiempo de caminata durante los 6 minutos del TM6M fue 455.8 ± 45.2 m.

En el grupo de AM dinapénicas, el 65 % es sedentaria y un 41 % presenta DMII, la edad promedio de este grupo fue 68.5 ± 3.6 años, el PCI promedio fue 97.1 ± 11.7cm, la FPM promedio de 22.7 ± 4.2 kg, la FCM4 promedio 115.3 ± 18.9 y la cantidad de metros promedio de caminata en el TM6M, fue 410.3 ± 81.1 m.

El 53 % de las mujeres obesas son sedentarias y el 29 % tienen DMII, la edad promedio de estas AM fue 67.7 ± 3.9 años, el PCI promedio fue 95.7 ± 11.5 cm, la FPM promedio de 23.6 ± 3.7 kg, FCM4 promedio de 115.9 ± 18.9 y el promedio de caminata, 428.6 ± 75.1 metros.

Finalmente, en el grupo de OD, el 66.7 % son sedentarias, aproximadamente 33 % tienen DMII, el promedio de PCI de estas mujeres es de 99 ± 10.8 cm, la FPM promedio fue 13.1 ± 5.8 kg, presentaron una FCM4 promedio de 115.6 ± 19.4 y un TM6M promedio de 413.5 ± 83.7 m. Se puede ver que estas últimas son las que presentan peores medidas de obesidad, dinapenia y TM6M.

Distribución Porcentual y medidas de resumen,
según Obesidad y Dinapenia

Variable.	G1	G2	G3	G4
Variables Biodemográficas				
Edad	68.50 ± 3.63	68.6 ± 1.8	67.7 ± 3.9	68.7 ± 3.5
Act.	17(35 %)	3(60 %)	35(40 %)	15(33,3 %)
Seden.	32(65 %)	2(40 %)	53(60 %)	30(66.7 %)
DMII	20(41 %)	1(20 %)	29(33 %)	15(33.3 %)
Variables Antropometricas y de Función Muscular				
Talla	149.9 ± 6.3	149.8 ± 4.1	150.5 ± 6.2	149.6 ± 6.2
Peso	68.1 ± 10.9	53.7 ± 4.7	68.2 ± 10.7	70 ± 10.4
IMC	30.4 ± 4.8	23.9 ± 1.5	30.2 ± 5.0	30.8 ± 4.5
PCA	103.7 ± 9.6	92.5 ± 2.7	103.8 ± 9.3	104.3 ±9.2
PCI	97.1 ± 11.7	77.8 ± 2.6	95.7 ± 11.5	99 ± 10.8
FPM	22.7 ± 4.2	23.8 ± 1.8	23.6 ± 3.7	23.1 ± 4.2
FIMV	13.5 ± 5.7	33.7 ± 7.3	23.3 ±12.2	13.1 ±5.8
TM6M	410.3 ± 81.1	455.8 ± 45.2	428.6 ± 75.1	413.5 ± 83.7
FCM4	115.3 ± 18.9	111 ± 25.5	115.9 ± 18.9	115.6 ± 19.4

Tabla 5.11: Medidas de resumen de AM y porcentaje según grupos (Obesas: PCI $< 80cm$; Dinapénicas: FIMV $<$ a $23.5kg$).

Una vez obtenida la predicción, se realizó una tabla de riesgo de padecer la enfermedad para los distintos niveles de DM con las variables seleccionadas por LASSO sin FCM4 (Figura 5.5), que resulta ser un importante aporte para una evaluación simplificada del riesgo de DM en mujeres adultas mayores activas de la ciudad de Valdivia, Chile. Y además, se realizó una segunda tabla de predicción para las variables seleccionadas por LASSO (Figura 5.6).

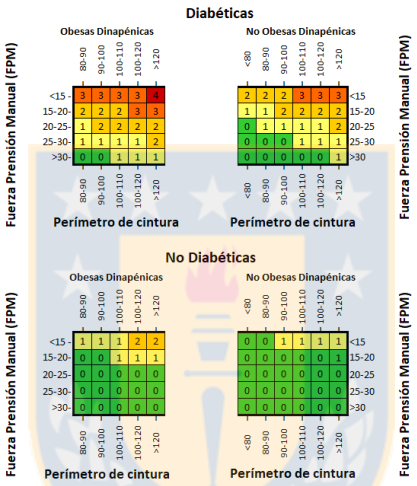

Figura 5.5: Tabla de predicción del riesgo de padecer disfunción motora en adultas mayores activas sin FCM4.

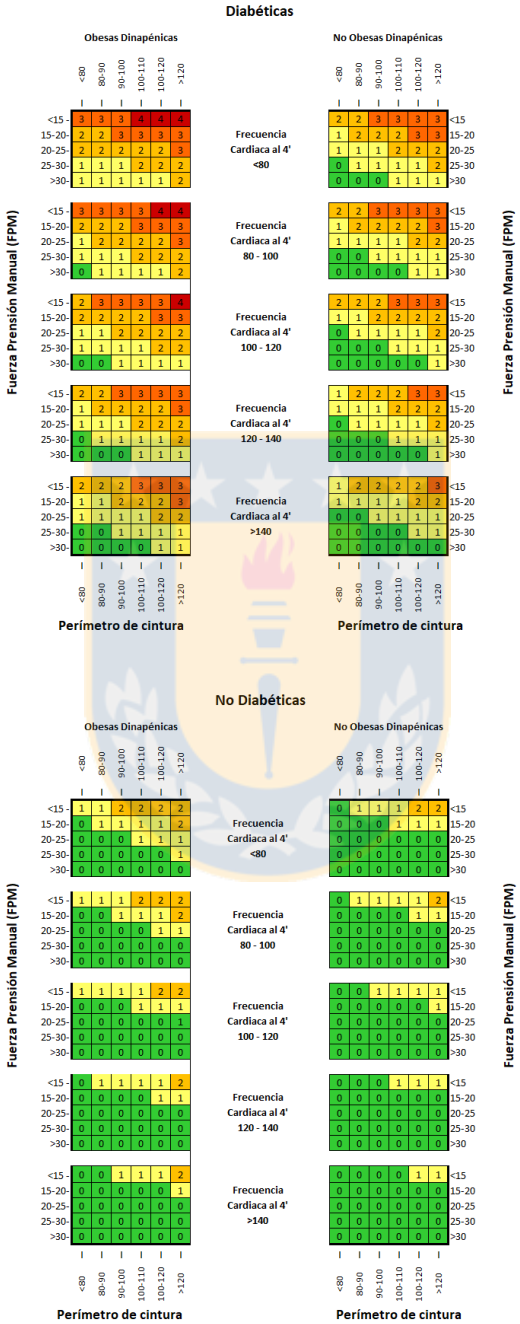

Figura 5.6: Tabla de predicción del riesgo de padecer disfunción motora en adultas mayores activas.

5.3. Discusión y Conclusión

5.3.1. Discusión

Al comparar la muestra de nuestro estudio con otros que se abocan a la misma problemática, resulta que, por una parte, los estudios presentados en Sénéchal (2012) y Bouchard et al. (2009), consideran mujeres con edades promedios de 65.4 y 70.1 años de edad, respectivamente; esta cifra es similar al promedio de edad de las participantes de nuestro estudio (67.52 años). Por otra parte, el 42 % de las participantes en nuestro estudio es sedentaria, lo que es bastante inferior a la tasa de 90 % nacional (MINSAL, 2017); esto se explica porque algunas de las participantes del grupo de estudio realizan actividad física en los centros de adulto mayor de la ciudad de Valdivia. Las participantes del estudio de Bouchard et al. (2009), incluyen un 51.2 % de sedentarias, porcentaje mayor al de nuestra investigación.

Los resultados de este estudio muestran que uno de los factores más significativamente asociados a la DM es la condición DMII, ésta medición unida a la información relativa a PCI, FPM, OD y FCM4 permiten simplificar el método de predicción considerando una caminata de 4 minutos.

La Society on Sarcopenia, Cachexia and Wasting Disorders concluyó, que un buen método para explicar la DM es el test de marcha en 6 minutos TM6M (Morley et al., 2011), medicion que requiere de un espacio físico de gran dimensión y que en nuestro estudio podría no ser utilizada.

Las variables que han sido seleccionadas en este estudio para la predicción de DM también están presentes en los hallazgos de los estudios antes mencionados. La FPM es una de las variables seleccionadas en nuestro estudio, lo que coincide con lo expuesto por Navia et al. (2012) y Morley et al. (2011), quienes mencionan que la pérdida de fuerza en el adulto mayor es un factor muy relevante cuando se habla de DM. Según Jenkins (2004) y el informe del MINSAL (2017), la Obesidad es un factor asociado a la DM. Sí bien esta variable no fue seleccionada en forma aislada para nuestro modelo predictivo, si se seleccionó el PCI, variable que define la Obesidad

para valores mayores a 80 cm en adultas mayores. Nuestro modelo también incluye la variable DMII como factor concominante asociado a DM, lo que es respaldado por los hallazgos del estudio de Muñoz (2016). Navia et al. (2012) y Bouchard et al. (2009) encontraron que las adultas mayores que se clasifican como OD tienen mayor grado de DM que las que no presentan esta condición. Precisamente, en esta investigación éste factor es una de las variables asociada a DM. Es así como los hallazgos de nuestro estudio se ven respaldados por la literatura relativa al tema, donde el aporte realizado consiste en afrecer una forma simplificada de diagnóstico de la DM y su nivel de gravedad.

5.3.2. Conclusión

El mayor aporte del presente estudio es la simplificación del diagnóstico de DM. En efecto, para poder diagnosticar esta condición en mujeres adultas mayores, se verifica que sólo es necesario determinar sí la paciente presenta DMII, registrar su PCI, FPM y verificar si se clasifica como OD. La ventaja de este hallazgo está en que la evaluación de DM podría ser realizada en una consulta médica de tamaño convencional. Además, se puede incluir la FCM4 en caso de tener la opción de aplicar el TM6M.

Si bien estos resultados solamente son extrapolables a mujeres adultas mayores activas de la ciudad de Valdivia, es posible replicar el estudio en otros contextos.

Otros aportes del estudio se relacionan con la importancia de la diabetes como factor de riesgo de la DM, concluyéndose que las mujeres adultas mayores Diabéticas y OD que presentan una FPM menor a 15 y una FC4M menor a 80 son las que tienen un mayor nivel de riesgo de padecer DM.

Finalmente, y más allá de los hallazgos mencionados, por una parte se provee a los profesionales de la salud de la ciudad de Valdivia de un instrumento de simple utilización para diagnóstico de DM y, por otra parte, se proporciona a los investigadores de otras regiones geográficas una metodología para generar modelos predictivos de DM.

Bibliografía

Acuña, E. (2008). Análisis de Regresión. Universidad de Puerto Rico.

Arahal, M. R., Soria, M. B. and Díaz, F. R. (2006). *Técnicas de Predicción con Aplicaciones en Ingeniería, 15*, 77-78. Universidad de Sevilla.

Arlot, S. and Celisse, A. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Statistics surveys, 4*, 40-79.

Bengio, Y. and Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of k-fold Cross-Validation. *The Journal of Machine Learning Research, 5*, 1089-1105.

Bouchard, D. R. and Janssen, I. (2009). Dynapenic-obesity and physical function in older adults. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, 65* (1), 71-77.

Breiman, L. (2017). Classification and regression trees. Routledge.

Bühlmann, P. and Van De Geer, S. (2011). Statistics For High-Dimensional Data: Methods, Theory and Applications. Springer Science & Business Media.

Cancino, B. and Navia, C. (2012). Disfunción Motora en Obesas Dinapénicas de Valdivia, Evaluadas entre Septiembre Y Diciembre de 2012. Tesis para optar al grado de Licenciado en Kinesiología. Escuela de Kinesiología, Universidad Austral, Chile.

Caner, M. and Fan, Q. (2010). The adaptive lasso method for instrumental variable selection. Working Paper, North Carolina State University.

Claerbout, J. F. and Muir, F. (1973). Robust Modeling With Erratic Data. *Geophysics, 38* (5), 826-844.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998). Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing, 20* (1), 33-61.

Daubechies, I., Defrise, M. and De Mol, C. (2004). An Iterative Thresholding Algorithm for Linear Inverse Problems With a Sparsity Constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 57* (11), 1413–1457.

Fan, J. and Li, R. (2001). Variable Selection Vía Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association, 96*, 456, 1348-1360.

Fielding, R. A., Vellas, B., Evans, W. J., Bhasin, S., Morley, J. E., Anne B. N. ... Mauro, Z. (2011). Sarcopenia: an undiagnosed condition in older adults. Current consensus definition: prevalence, etiology, and consequences. International working group on sarcopenia. *Journal of the American Medical Directors Association, 12* (4), 249-256.

Frank, L. E and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics, 35* (2), 109-135.

Friedman, J. H. (1979). A tree-structured approach to nonparametric multiple regression. Springer, 5-22.

Fu, W. J. (1998). Penalized Regressions: The Bridge Versus The LASSO. *Journal of Computational and Graphical Statistics, 7* (3), 397-416.

Hernández, P. A. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista colombiana de estadística, 27* (2), 139-151.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning, 112. New York: springer.

Jenkins, K. R. (2004). Obesity's effects on the onset of functional impairment among older adults. *The Gerontologist, 44* (2), 206-216.

Knight, K. and Fu, W. (2000). Asymptotics for LASSO-Type Estimator. *Annals of Statistics, 28* (5), 1356-1378.

Lebart, L., Morineau, A. and Piron, M. (1995). Statistique exploratoire multidimensionnelle. Paris: Dunod.

Melzer, D. and Parahyba, M. I. (2004). Socio-demographic correlates of mobility disability in older Brazilians: results of the first national survey. *Age and Ageing, 33* (3), 253–259.

MINSAL, Pontificia Universidad Católica de Chile, Universidad Alberto Hurtado MINSAL. Encuesta Nacional de Salud ENS Chile 2009-2010, Chile.

MINSAL, Pontificia Universidad Católica de Chile, Centro UC. Encuesta Nacional de Salud ENS Chile 2016-2017, Chile.

Morley, J. E., Abbatecola, A. M., Argiles, J. M., Baracos, V., Bauer, J., Bhasin, S., and Fearon, K. (2011). Sarcopenia with limited mobility: an international consensus. *Journal of the American Medical Directors Association, 12* (6), 403-409.

Muñoz, G., Degen, C., Schröder, J., and Toro, P. (2016). Diabetes Mellitus y su asociación con deterioro cognitivo y demencia. *Revista Médica Clínica Las Condes, 27* (2), 266-270.

NCSS. (2018). Stepwise Regression. *NCSS Statistical Software, 311* (1), 1-2.

Orea, S. V., Vargas, A. S. and Alonso, M. G. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene, 779* (73), 33.

Park, C. and Yoon, Y. J. (2011). Bridge Regression: Adaptivity and Group Selection. *Journal of Statistical Planning and Inference, 141* (11), 3506-3519.

Pötscher, B. M. and Leeb, H. (1997). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis, 100* (9), 2065-2082.

Reid, S., Tibshirani, R. and Friedman, J. (2013). A Study of Error Variance Estimation in LASSO Regression.

Santosa, F. and Symes, W. W. (1986). Linear Inversion of Band-Limited Reflection Seismograms. *SIAM Journal on Scientific and Statistical Computing, 7* (4), 1307-1330.

Sénéchal, M., Dionne, I. J. and Brochu, M. (2012). Dynapenic abdominal obesity and metabolic risk factors in adults 50 years of age and older. *Journal of aging and health, 24* (5), 812-826.

Sill, M., Hielscher, T., Becker, N. and Zucknick, M. (2014). c060: Extended Inference With LASSO and Elastic-Net Regularized Cox and Generalized Linear Models. *Journal of Statistical Software, 62* (5), 1-22.

Tibshirani, R. (1997). The LASSO Method for Variable Selection In The Cox Model. *Statistics in Medicine, 16* (4), 386-395.

Tibshirani, R.(1996). Regression Shrinkage and Selection Via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological), 58* (1), 267-288.

Taylor, H. L., Banks, S. C. and McCoy, J. F. (1979). Deconvolution with the L 1 norm. *Geophysics, 44* (1), 39-52.

Wu, T. T. and Lange, K. (2008). Coordinate Descent Algorithms for LASSO Penalized Regression. *The Annals of Applied Statistics, 2* (1), 224-244.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009). Genome-Wide Association Analysis by LASSO Penalized Logistic Regression. *Bioinformatics, 25* (6), 714-721.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic-Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67* (2), 301-320.