Universidad de Chile
Instituto de Salud Poblacional
Programa de Magíster en Bioestadística

Determinación del Sexo a través de Variables Métricas de Clavícula en Osamentas Humanas: Aplicación de técnicas de regresión Logit y MARS.

Tesis para optar al Grado Académico de

Magíster en Bioestadística

Alumna: Paulina Marambio Vásquez

Profesor Guía:
Sergio Alvarado Orellana

Santiago, Diciembre 2017

Índice

Página

Resumen

En Antropología Forense la reconstrucción del perfil biológico es fundamental para la identificación de los fallecidos. Dentro de ésta, la determinación del sexo constituye una de las labores más importantes del análisis pericial, siendo posible de estimar en base a la morfología esqueletal o la dimensión de huesos largos. Esta última estrategia es útil cuando no se encuentran regiones diagnósticas (pelvis y cráneo) y/o cuando existen piezas aisladas del esqueleto. La clavícula pieza doble de la cintura escapular ha sido ampliamente utilizada para estimar el sexo en individuos no identificados. En este estudio se aplicó análisis de regresión logística y la metodología MARS (Multivariate adaptive regression splines) para generar modelos predictivos del sexo a partir de una muestra de entrenamiento (n=85 individuos, 52 masculinos y 33 femeninos), los cuales fueron evaluados en una muestra de validación compuesta por n=37 individuos (24 masculinos y 13 femeninos). Los modelos generados fueron comparados en cuanto a indicadores de bondad, tales como sensibilidad y especificidad, registrándose para la regresión logística un 95% para el primer indicador y un 88.88% para el segundo, con una cifra de 93.1% de los casos correctamente clasificados, mientras que para el análisis MARS, se observó un 41.67% de sensibilidad y un 92.31% en cuanto a especificidad y 59.46% de los casos correctamente clasificados, posicionando a la regresión logística como una herramienta útil para la discriminación del sexo en osamentas no identificadas, de fácil aplicación e interpretación.

Capítulo I: Introducción

En Antropología Física, el análisis de restos humanos esqueletizados de un fallecido es una de las principales labores que proporciona información referente al individuo, permitiendo reconstruir el perfil biológico, los patrones de subsistencia y en general, el modo de vida de los individuos examinados [1]. Otra perspectiva, pero basada en el mismo objetivo la tiene la Antropología Forense, que además de entregar información acerca de características morfológicas de los sujetos, indaga en las lesiones que el individuo exhibe en los huesos y que podrían ser el reflejo de acciones de terceros y, que junto con otros elementos como el contexto del hallazgo del cuerpo, otorgaría relevancia médico legal al caso estudiado, haciendo que los organismos judiciales ejerzan su labor penal [2].

Las herramientas que el análisis de osamentas aplica para la identificación humana consideran variables de diverso tipo [3], como dicotómicas o binarias (e.g.: sexo), categóricas (e.g.: ancestría) y continuas (e.g.: estatura, edad, etc.)

La estimación de sexo es uno de los procedimientos más relevantes en la identificación de osamentas humanas. Las metodologías para llevar a cabo esta tarea implican el examen macroscópico de la morfología de las regiones más dimórficas del esqueleto humano (cráneo y pelvis) [4], la caracterización de las regiones de interés en los restos analizados y, la posterior clasificación del individuo en femenino o masculino. Sin embargo, cuando el esqueleto está incompleto o el material de análisis se conforma por restos óseos mezclados de más de un sujeto, las metodologías basadas en el examen de variables métricas de huesos largos son fundamentales en la estimación del sexo de los individuos [5].

Durante el desarrollo de la disciplina antropológica, distintas técnicas se han implementado con el propósito de conocer esta característica en los individuos analizados, distinguiéndose de acuerdo a los hitos anatómicos considerados para el diagnóstico del sexo o el tipo de variable estudiada en los restos óseos. El esqueleto adulto muestra el efecto de procesos fisiológicos que generan diferencias entre los individuos masculinos y femeninos lo cual se conoce como dimorfismo sexual [1,3].

La clavícula, pieza doble que conforma la cintura escapular del esqueleto ha sido útil en el diagnóstico del sexo. Para estimar esta característica, se han utilizado de manera general, dos dimensiones de la pieza: su largo máximo y el contorno del eje medio [6]. De manera previa, otros autores han confirmado que las variables métricas de clavícula son buenos predictores del sexo en osamentas [6, 7, 8].

Las diversas poblaciones humanas se han adaptado a distintos medios geográficos y culturales generando una gran variabilidad inter poblacional. Esta condición obliga a que las distintas técnicas de reconstrucción del perfil biológico de los individuos sometidos a análisis sean evaluadas y ajustadas al grupo al que se aplican.

El principal objetivo de esta tesis es construir un modelo estadístico que permita predecir el sexo en osamentas de individuos no identificados, usando variables asociadas a la clavícula, tales como: longitud máxima y circunferencia de su eje medio.

Capítulo II: Marco Teórico

2.1 Marco Teórico de Referencia

En Antropología Forense la reconstrucción del perfil biológico es fundamental en la identificación del individuo y busca conocer características como el sexo, la edad, la estatura y el patrón ancestral. Las metodologías que permiten estimar el sexo en el esqueleto se clasifican en aquellas que consideran variables morfológicas y en otras que basan su conclusión en el examen de variables métricas [3]. Para el primer caso, las regiones esqueletales que son examinadas son el cráneo y la pelvis, siendo esta última, la más diagnóstica del sexo, y para la segunda situación, son analizados, de preferencia los huesos largos del esqueleto apendicular [9], considerando por ejemplo, su longitud.

Expertos en Antropología Física han desarrollado herramientas de identificación en osamentas, a lo largo de la historia de la disciplina, formulando estándares en base a muestras de poblaciones específicas. Sin embargo, la enorme variabilidad de los grupos humanos, hace necesaria la evaluación de estas herramientas metodológicas, puesto que un método no garantiza la correcta estimación de una característica en un individuo que pertenece a otro grupo [8]. Es por esta razón que los equipos de antropólogos son motivados por la necesidad de evaluar los métodos ya establecidos o buscar nuevos estándares que resuelvan el problema de la identificación en osamentas.

Con la aparición de las características sexuales secundarias, producto de la secreción hormonal diferenciada entre hombres y mujeres, el esqueleto de los individuos experimenta una serie de modificaciones, denominándose este conjunto de cambios como dimorfismo sexual. Las diferencias entre esqueletos femeninos y masculinos se resumen en distinciones de tamaño y forma, siendo los individuos masculinos de mayor tamaño y robustez (inserciones musculares marcadas), mientras que los esqueletos femeninos se caracterizan por ser pequeños y gráciles [9,10]. Estos criterios son usados al momento de determinar el sexo en osamentas de individuos de identidad desconocida. En

clavícula, el dimorfismo sexual se manifiesta por mayor curvatura y rugosidad de inserciones musculares, alto grado de robustez y mayor tamaño de las piezas en individuos masculinos [8]. Sin embargo, cabe señalar que la robustez ósea (en parte reflejada por el aspecto de las inserciones muscular) es una característica que podría también estar influida por factores externos tales como los patrones de actividad desarrollados por el individuo (p. ej. Estrés mecánico producido por la actividad física). En otros términos:

> "…algunos rasgos varían en función del ambiente en el que viven o han vivido los individuos. Por ejemplo, aquellos que han sufrido algún tipo de estrés mecánico en una o varias partes del cuerpo por realizar cotidianamente una tarea específica tienen, comparados con individuos que no realizaron la misma actividad, diferencias morfológicas que se evidencian en los restos óseos. Por eso, su estudio permite inferir los estilos de vida que esas poblaciones tuvieron en el pasado" [11].

Los patrones de actividad que son parte de la cultura de los individuos influyen en la morfología esqueletal, al respecto Kottak señala:

"La cultura es una fuerza medioambiental clave que determina cómo crecen y se desarrollan los seres humanos" [12]. Es preciso, por lo tanto, considerar la morfología ósea con precaución, en términos de que se requiere conocer o al menos tener una idea general de las actividades que realizaron los individuos analizados.

Sin embargo, frente a las afirmaciones mencionadas arriba, cabe reflexionar sobre el concepto de sexo como una construcción social que limita las posibilidades a solo dos alternativas: hombre o mujer, sin considerar el rango de variación en la expresión de las características del sexo tanto en el nivel genotípico como en el fenotípico. Es así como

esta clasificación es limitante en casos de individuos con anomalías de tipo cromosómico, gonadal, endocrino o fenotípico [10].

Ya en el siglo XX, la clavícula es considerada una pieza importante para el diagnóstico del sexo en osamentas. En 1966 Jit y Singh [6] reunieron una muestra de n=122 individuos de ambos sexos, pertenecientes a población punjabee, a partir de la cual se establecieron criterios para discriminar sexo en base a dimensiones de clavícula (largo máximo, circunferencia del eje medio de la pieza y peso de la misma) sometiendo estas características a un análisis univariado y del cual, se obtuvieron puntos de demarcación para discriminar a individuos masculinos de los femeninos.

La posterior formulación de estándares de clasificación en base a medidas de clavícula se diversificó: se consideraron las variables separadamente o de manera conjunta, se trabajó solo con clavícula o se agregaron otras piezas óseas.

2.2 Marco Teórico Estadístico

El análisis de función discriminante lineal ha sido usado extensamente, entregando una buena precisión en la generación de estándares para el diagnóstico del sexo. Otros estudios han sometido sus datos a análisis de componentes principales además de discriminante [7,13] y también se ha probado la eficacia del análisis de regresión logística. Al respecto, este análisis es más robusto y trabaja bien e incluso mejor, con menos supuestos estadísticos que el análisis de función discriminante [5].

En población centroamericana, específicamente en guatemaltecos se estableció el sexamiento métrico para diversas piezas óseas entre las cuales se incluye la clavícula [9]. Usando análisis de función discriminante, se determinaron los siguientes criterios:

Para longitud máxima se tiene la siguiente fórmula:

$$1.197 * X - 16.738 \quad , (1)$$

donde *X* corresponde al largo de la clavícula en cm, con un valor crítico de – 0.3175 (punto de corte).

Los individuos menores a este valor son femeninos y aquellos mayores, son masculinos con una probabilidad del 88.8%, mientras que para la circunferencia del eje medio, la fórmula establecida es:

$$3.434 * X - 11.197 \quad , \quad (2)$$

, con valor crítico de – 0.2350 (clavículas con circunferencia menor a este valor corresponden a femeninas y mayor a este punto de corte, masculinas con una probabilidad de 86.1%).

Las poblaciones latinoamericanas han sido las menos estudiadas con respecto al establecimiento de estándares de sexamiento, sin embargo, merecen mayor atención, considerando las violaciones a los derechos humanos cometidas durante gobiernos de dictadura [14]. Para estos casos, ya el año 2014, se efectuó un estudio que buscó definir criterios para determinar el sexo en base a dimensiones de varias piezas óseas (incluída clavícula y considerando solo su largo máximo), usando una muestra de individuos chilenos provenientes de la Colección Osteológica Juan Munizaga (individuos cuya data de muerte se sitúa entre los años 1950 y 1970) y aplicando análisis de función discriminante. Sin embargo, la validación de esta herramienta aún está en curso [15].

Se requiere el perfeccionamiento de los métodos en construcción y además, contar con criterios que provengan del estudio de muestras de poblaciones más recientes que puedan ser aplicados en la estimación del sexo de casos forenses con data más actual. Por esta razón y para esta investigación, la muestra que se empleará la conforman un conjunto de individuos fallecidos en la década de los noventa.

En el ámbito de la construcción de criterios para establecer el sexo en osamentas y como ya se mencionó, el análisis de función discriminante ha sido ampliamente utilizado. Sin embargo, existen otras alternativas como el análisis de regresión logística (logit), que vale la pena evaluar a la hora de medir el ajuste de variables consideradas en un modelo adecuado para el problema estudiado. Ésta se caracteriza por ser más robusta y mejor que el análisis discriminante, puesto que debe cumplir con menos supuestos estadísticos. Además, el modelo logit calcula un p-value con el cual, por una parte, se clasifica un individuo desconocido y por otro lado, sirve para efectuar una declaración de probabilidad sobre la verosimilitud de una correcta estimación para el caso dado [5]. Forma parte de los métodos de regresión para variables dependientes binarias, compartiendo características con otros modelos de regresión para variables dependientes categóricas [16]. Corresponde al modelo de regresión binaria o de elección discreta en que las alternativas de opción se reducen a dos posibilidades mutuamente excluyentes (o dummy), codificadas como 0 para una respuesta negativa (el evento no ocurre) o 1 para una respuesta positiva (el evento ocurre). Esta característica, clasifica a la técnica a emplear en lo que se conoce como modelos de probabilidad no lineales, cuya función de especificación utilizada garantiza un resultado en la estimación comprendido en el rango 0 a 1. La función de distribución del modelo logit, corresponde a la distribución logística [17].

Por su parte el modelo MARS (Multivariate Adaptive Regression Splines), es una herramienta altamente automatizada[1], que permite efectuar análisis de regresión lineal escalonada, pero que puede modelar no-linealidades e interacciones entre variables. Es un método de regresión flexible para una gran dimensión de datos, que toma la forma de una expansión en producto de funciones base en ranura (una función polinomial definida en tramos, que es suave y donde las piezas polinomiales se conectan mediante suavizamientos), donde el número de funciones base, así como los parámetros asociados con cada uno están automáticamente determinadas por los datos, lo cual está basado en el enfoque de particionamiento recursivo para la regresión. El modelo logra identificar de manera separada las contribuciones aditivas y aquellas asociadas con las diferentes interacciones multivariables [18, 19, 20]. Es una metodología aplicable tanto a variables dependientes o respuesta de tipo continuas como binarias.

El modelo MARS tiene la forma general:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m \, h_m(X) \quad , (3)$$

- $h_m(X)$ es una función llamada “basal” y corresponde a una expresión en donde se incorpora un valor conocido como “nodo”, el cual particiona el rango de la variable predictora en subregiones las que posteriormente mediante un algoritmo se seleccionan para construir el modelo final.

- Betas son los coeficientes estimados al minimizar la suma residual de cuadrados.

- Estos coeficientes pueden ser considerados como pesos que representan la importancia de la variable [18].

[1] https://www.salford-systems.com/videos/training/mars

El análisis de regresión logística y el modelo MARS detallados más arriba, son las dos técnicas que se aplicarán en este trabajo de investigación.

Los criterios que determinaron la selección de los métodos escogidos, por una parte, consisten en que ambas técnicas son novedosas en el ámbito de la estimación del sexo en restos óseos, corresponden a modelos no paramétricos, los cuales se eximen del cumplimiento de supuestos estadísticos (normalidad, homocedasticidad, etc...). Esta alternativa no paramétrica es absolutamente adecuada cuando existen datos *missing* ya que la técnica construye funciones basales que penalizan aquellas variables que presentan valores faltantes. Por otra parte, el análisis de regresión logística o modelo Logit es utilizado para predecir el resultado de una variable categórica en función de las variables independientes. Mientras que el modelo MARS se caracteriza por ser flexible en cuanto al tipo de variables respuesta y predictoras que forman parte del modelo.

2.2.1. Regresión Logística

De acuerdo a lo expuesto por Hosmer y Lemeshow [21] como en cualquier modelo de regresión, la regresión logística busca describir la relación entre una variable respuesta y una o más variables explicativas, hallando el mejor ajuste entre éstas. La variable respuesta es de tipo dicotómica o binaria.

Dado un conjunto de *p* variables independientes denotadas por el vector $x' = (x_1, x_2, \dots, x_p)$ y siendo la probabilidad condicional de que un evento ocurra (la respuesta esté presente), $P\,(Y = 1|x) = \pi(x)$ el logit del modelo de regresión logística múltiple está definido por la siguiente ecuación:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad , (4)$$

siendo el modelo de regresión logística:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad , (5)$$

De acuerdo a lo expuesto por Escobar *et al* [22] el modelo de regresión logística se estima mediante el método de máxima verosimilitud, obteniendo los valores de los parámetros *β* que con mayor probabilidad pueden haber generado los valores de la variable dependiente de la muestra. Mediante un proceso iterativo se prueban distintos valores para los parámetros *β* hasta que se encuentran aquellos valores que maximizan la función de verosimilitud.

Una vez obtenidos los coeficientes de la regresión la ecuación estimada sería

$$ln\frac{\Pr\,(y = 1)}{\Pr\,(y \neq 1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad , (6)$$

En la regresión logística es difícil interpretar los coeficientes directamente como en la regresión lineal, por lo que se debe transformar la ecuación en una más fácil de interpretar, que logre mostrar de manera comprensible, la relación entre las variables independientes y dependiente del modelo.

Existen dos transformaciones de la ecuación logit que posibilitan su inmediata interpretación. Una de éstas elimina el logaritmo situado en el lado derecho de la ecuación original de la siguiente forma:

$$\Omega(\mathrm{x}) = \frac{\Pr(y = 1|x)}{1 - \Pr(y = 1|x)} = \exp\,(b_0 + b_1 x_1 + \cdots + b_k x_k) \quad , (7)$$

así, los coeficientes indican cómo varía la razón de ocurrencia de la variable dependiente, cuando cambia en una unidad el valor de las variables independientes. Sin embargo,

sigue siendo esta una forma compleja de interpretar una regresión logística, puesto que el cambio en la variable dependiente es expresado en forma de cociente de razones.

Una segunda transformación muestra en el lado derecho de la ecuación variaciones en la probabilidad de ocurrencia del suceso:

$$\Pr(y = 1|x) = \frac{\exp(b_0 + b_1 x_1 + \cdots + b_k x_k)}{1 + \exp(b_0 + b_1 x_1 + \cdots + b_k x_k)} \quad , (8)$$

Sin embargo, esta corresponde a una relación no lineal. Una alternativa que modifica completamente la estrategia de análisis es estudiar la ecuación logística en forma probabilística, usando las probabilidades predichas por el modelo para valores puntuales de las variables independientes.

Bondad de ajuste del modelo Logit

De acuerdo a lo expuesto por Medina [17], evaluar el ajuste del modelo es relevante por dos razones: identificar problemas en el modelo, por causa de datos erróneos o una por una mala especificación de las variables y para valorar su capacidad explicativa

El ajuste del modelo de regresión logística puede evaluarse con diversas técnicas:

- Índice de cociente de verosimilitudes
- Estadístico Ji-Cuadrado de Pearson.
- Porcentaje de aciertos estimados en el modelo

Índice de cociente de verosimilitudes

Evalúa el ajuste del modelo comparando el valor de la función de verosimilitud de dos modelos, el modelo completo (modelo estimado con todas las variables explicativas) y el restringido (modelo en el que la constante es la única variable explicativa) de la siguiente forma:

$$RV = ICV = 1 - \frac{log\ L}{log\ L(0)} \quad , (9)$$

Siendo L el valor de la función de verosimilitud del modelo completo y $L(0)$ el valor del modelo restringido.

El resultado de este índice fluctúa entre los valores 0 y 1. Aquellos cercanos a 0 se obtienen cuando $L(0)$ sea muy similar a L, esto ocurre cuando las variables del modelo son poco significativas y por ende la capacidad explicativa del modelo es reducida. Por el contrario, mientras más cercano a 1 sea el valor del índice calculado, mayor es la capacidad explicativa del modelo.

Estadístico Ji-cuadrado de Pearson

Es una medida del error que cuantifica la diferencia entre el valor observado y el estimado de la siguiente manera:

$H_0: Y_i = \widehat{M}_i$; lo que equivale a $H_0: Y_i - \widehat{M}_i = e_i = 0$ ($\widehat{M}_i$es la probabilidad de que el evento ocurra).

Se calculan los residuos estandarizados de Pearson del modelo (diferencia entre el valor observado de la variable respuesta y el estimado, dividido por la estimación de la desviación típica, ya que la esperanza es nula:

$$\frac{Y_i - \widehat{M}_i}{\sqrt{\widehat{M}_i(1 - \widehat{M}_i)}} \quad , (10)$$

Luego y mediante el contraste de multiplicadores de Lagrange se calcula el Ji-cuadrado de Pearson:

$$\chi^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{M}_i)^2}{\widehat{M}_i(1 - \widehat{M}_i)} \quad , (11)$$

Mientras más cercano a cero se encuentre el estadístico, mejor será el ajuste del modelo. Se requiere saber la distribución de este estadístico para conocer a partir de qué valor el ajuste es aceptable. Bajo la hipótesis de nulidad, éste se distribuye como una Ji-cuadrado con (*n-k*) grados de libertad, su valor se compara con el valor teórico de la tabla de Ji-cuadrado para contrastar la hipótesis de nulidad. Si éste valor es mayor al valor teórico, se rechaza la hipótesis de nulidad, es decir el error cometido es significativamente distinto de cero (mal ajuste).

Porcentaje de aciertos estimados en el modelo

Esta estrategia de evaluación se basa en predecir con el modelo los valores de la variable respuesta Y_i de tal forma que $Y_i = 1$ si $\widehat{M}_i > c$ o $Y_i = 0$ si $\widehat{M}_i < c$, donde c corresponde a un valor umbral que permite determinar si el valor de la predicción es 1 o 0, siendo este umbral generalmente igual a 0.5, lo cual no es la alternativa más conveniente en el caso de que la muestra tenga desequilibrios entre la cantidad de unos y ceros, puesto que llevaría en a no predecir ningún uno o ningún cero. La solución a este inconveniente es considerar un umbral inferior.

Cualquier tipo de regla predictiva elegida conducirá a cometer dos errores, se determinarán erróneamente ceros como unos y unos se clasificarán incorrectamente como ceros.

Al disminuir el valor umbral por debajo de 0.5, aumentarán las veces en la que se clasifican correctamente observaciones como unos para las que $Y_i = 1$, pero también se incrementará el número de veces en que se clasifiquen observaciones como unos cuando $Y_i = 0$. En definitiva, el cambio en el valor umbral implica la disminución de la probabilidad de un tipo de error y el aumento de la probabilidad del otro tipo de error. Así, el valor del umbral dependerá de la distribución de los datos en la muestra y de la importancia relativa de cada tipo de error.

Cuando ya se ha elegido el valor umbral y dado que los valores reales de Y_i se conocen, se contabiliza el porcentaje de aciertos para establecer si la bondad de ajuste es elevada o no.

A partir de lo anterior, es posible obtener la siguiente tabla de clasificación de aciertos (Tabla 1):

Tabla 1. Clasificación de aciertos.

		Valor real de Y_i	
		$Y_i = 0$	$Y_i = 1$
Predicción de $\widehat{M}_i$	$\widehat{M}_i < c$	P_{11}	P_{12}
	$\widehat{M}_i > c$	P_{21}	P_{22}

Siendo P_{11} y P_{22} predicciones correctas (valores 0 y 1 correctamente predichos, respectivamente) y P_{12} P_{21} predicciones erróneas (valores 1 mal predicho y valores 0 mal predichos, respectivamente).

De acuerdo los valores definidos anteriormente, es posible definir seis índices para evaluar la bondad del ajuste:

- **Tasa de aciertos:** que es el cociente entre predicciones correctas y el total de predicciones ($\frac{P_{11}+P_{22}}{P_{11}+P_{12}+P_{21}+P_{22}}$)

- **Tasa de errores:** corresponde al cociente entre predicciones incorrectas y el total de predicciones ($\frac{P_{12}+P_{21}}{P_{11}+P_{12}+P_{21}+P_{22}}$)

- **Especificidad:** se define como la proporción entre la frecuencia de valores 0 correctos y el total de valores 0 observados ($\frac{P_{11}}{P_{11}+P_{21}}$).

- **Sensibilidad:** corresponde a la razón entre los valores 1 correctos y el total de valores 1 observados ($\frac{P_{22}}{P_{12}+P_{22}}$).

- **Tasa de falsos ceros:** es la proporción entre la frecuencia de valores 0 incorrectos y el total de valores 0 observados ($\frac{P_{21}}{P_{11}+P_{21}}$).

- **Tasa de falsos unos:** es la razón entre los valores 1 incorrectos y el total de valores 1 observados ($\frac{P_{12}}{P_{12}+P_{22}}$).

2.2.2. MARS (Multivariate Adaptive Regression Splines)

Con el objeto de definir esta técnica es pertinente describir cada uno de los conceptos de la sigla MARS:

- Multivariate (multivariado): esta técnica es capaz de generar un modelo basado en un conjunto de variables de entrada, independientes o predictoras.

- Adaptive (adaptado): produce modelos flexibles en distintos pasos cada vez que se ajusta el modelo.

- Regression (regresión): Estima la relación entre variables independientes y dependientes.

- Spline (ranura): produce una función polinomial definida por una pieza o trozo, que se caracteriza por ser suave (al poseer derivadas de tercer orden) y que se conecta con otras funciones polinomiales a través de nodos (knot).

El modelo MARS corresponde a una generalización del método de Regresión recursiva particionada que toma la forma de una expansión en producto de funciones base spline, donde la cantidad de éstas y la ubicación de los nodos, son definidos automáticamente por los datos. Es capaz de modelar relaciones aditivas o interacciones entre variables e identificar contribuciones aditivas y éstas con las diferentes interacciones multivariadas [18].

De acuerdo a lo expuesto por Oyarzún [20], la regresión recursiva particionada es una aproximación a la función desconocida $f(x)$ en x usando una expansión en un conjunto de funciones base, donde cada una de éstas es producto de funciones de salto univariante. La explicación de la metodología MARS se detallará a continuación en base a los trabajos de Friedman [18, 23, 24, 25].

Si se tienen *N* observaciones de las variables $y = (x_1, x_2, \dots x_N)$ y $x = (x_1, x_2, \dots x_N)$ definidas como $\{y_i, x_i\}_{i=1}^{N}$. Sea $\{R_j\}_{j=1}^{S}$ un conjunto de subregiones disjuntas de D, tal que $D = \cup_{j=1}^{S} R_j$ Dadas las subregiones $\{R_j\}_{j=1}^{S}$ la partición recursiva estima la función $f(x)$ en x como:

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(x) \qquad , (12)$$

B_m es una función base de la forma:

$$B_m(x) = I[\, x \in R_m] \qquad , (13)$$

$I[.]$: Función indicadora con valor 1 si el argumento es verdadero y 0 en otro caso.

$\{a_m\}_1^M$: son los coeficientes de la expansión fijados conjuntamente para un mejor ajuste de los datos. Los $\{R_m\}_1^M$ corresponden a las subregiones del espacio de las covariables. *H* [η] es una función indicadora, que es un producto de funciones de salto univariante,

$$H[\eta] = \begin{cases} 1, & si\ \eta \geq 0, \\ 0, & en\ otro\ caso \end{cases} \qquad , (14)$$

Lo anterior describe cada subregión R_m, siendo $B_m(x)$ una función con valores 1 si y sólo si x es un miembro de los $R_m th$ subregión de *D.*

Por su parte la regresión spline es desarrollada como estrategia para enfrentar las limitaciones formales de la Partición recursiva y lo hace estimando la función desconocida $f(x)$, a partir de la función spline que se obtiene al dividir el rango de la variable predictora x en K+1 regiones disjuntas separadas por K nodos. Este resultado tiene forma de polinomio de grado q en cada una de las subregiones obtenidas por la partición, para que la función y sus primeras $q-1$ derivadas sean continuas en cada región. El resultado de esta estrategia son funciones ajustadas y suavizadas.

Para la función spline se considera un orden menor o igual que tres, esto porque cada polinomio de grado $q-1$ parámetros y como se tienen $K+1$ regiones, se deben estimar $(K+1)(q+1)$ parámetros por el método de mínimos cuadrados.

Por continuidad en cada región se van agregando q restricciones para cada nodo, teniéndose como resultado $K\ q$ restricciones, para luego estimar $K+\ q+1$ parámetros.

La regresión spline se ajusta eligiendo un conjunto de funciones base $\{B_k^{(q)}(x)\}_0^{K+q}$ que forma el espacio de las funciones spline de orden q, aplicando a su vez el método de mínimos cuadrados.

Según lo anterior, la aproximación se formaliza como $\hat{f}(x) = \sum_{k=0}^{K+q} a_k\, B_k^{(q)}(x)$ en donde $\{a_k\}_0^{K+1}$ corresponden a los coeficientes sin restricciones, las cuales se incluyen en las funciones base $\{B_k^{(q)}(x)\}_0^{K+q}$, la que a su vez se compone de funciones:

$$\{x^f\}_0^q, \{(x-t_k)_+^q\}_1^K \qquad \text{, (15)}$$

Por su parte, $(x-t_k)_+^q = \begin{cases} 0 & , x \le t_k \\ (x-t_k)^q, & x \ge t_k \end{cases}$

determina la ubicación de los nodos que definen las $K+1$ regiones y las funciones de dominio truncadas.

Los nodos marcan los límites de una región de datos, describiendo el comportamiento de los cambios en la función. En regresión spline estas estructuras son predeterminadas e igualmente espaciadas, mientras que en MARS, son escogidos por una estrategia de búsqueda automática, que implica una minimización de la Validación Cruzada Generalizada (sigla en inglés GCV):

$$\sum_{i=1}^{N}(y_i - \sum_{k=0}^{K+q} a_q B_k^{(q)}(x))^2 \quad , (16)$$

Se sugiere el uso de funciones basales truncadas que al reemplazarse en la ecuación anterior resulta:

$$\sum_{i=1}^{N}\{y_i - \sum_{j=0}^{q} b_j x^j - \sum_{k=1}^{K} a_k (x-t_k)_+^q\}^2 \quad , (17)$$

Los coeficientes $\{b_j\}_0^q$, $\{a_k\}_1^K$ corresponderían a los parámetros asociados con una regresión múltiple de respuesta y, sobre los predictores $\{x^j\}_0^q$ y $\{(x-t_k)_+^q\}^K$ respectivamente.

Agregar o quitar un nodo, significa sumar o eliminar la variable respectiva $(x-t_k)_+^q$, puesto que se elige automáticamente el nodo y su localización.

Se comienza con una gran serie de nodos $(t_1, \ldots., t_{K_{max}})$ donde $K_{max} = N-2$ y las correspondientes variables $\{(x-t_k)_+^q\}^{K_{max}}$ que son candidatas a ser seleccionadas por el método stepwise. Las observaciones anómalas afectan las respuestas solo localmente.

En casos multivariados, es decir para p variables $x = \{x_1, \dots, x_p\}$, el espacio dimensional es R^P dividido en un grupo de regiones disjuntas y en cada una, $\widehat{f_q}(x)$ es un polinomio de p variables. $\widehat{f_q}(x)$ se restringe a una región disjunta y para que todas sus derivadas de orden $q-1$ sean continuas en todas sus partes. Así se generan restricciones en las regiones disjuntas y en los límites de los polinomios. La aproximación spline se efectúa eligiendo un conjunto de funciones base que generen el espacio de las funciones spline.

Cuando $p > 2$ se toman regiones disjuntas que definen la aproximación spline como productos tensores de intervalos disjuntos en cada una de las variables delineadas por la ubicación del nodo. De esta forma se tiene K_j nodos en cada una de las variables $(0 \leq j \leq p)$ generando $\prod_{j=1}^{p} (K_{j+1})$ regiones. El producto tensorial de las correspondientes funciones basales spline unidimensionales asociadas a la ubicación de los nodos de cada variable corresponde al grupo de funciones base que forman el espacio de funciones spline sobre todo el conjunto de regiones.

$$\hat{f}\,(x) = \sum_{k_{1=0}}^{K_1+q} \dots\dots \sum_{k_{n=0}}^{K_p+q} a_{k_1 \dots k_p} \prod_{j=1}^{p} B_{k_j}^{(q)}\,(x_j) \qquad , (18)$$

Donde $\{B_{k_j}^{(q)}\,(x_j)\}_{K_{j=0}}^{K_j}$ corresponde al conjunto de funciones base para la aproximación spline de orden q dada la ubicación de los nodos K_j en x_j. El término $\prod_{j=1}^{p}(\,K_j + q + 1)$ es el número de coeficientes para ser estimados.

MARS es la generalización de la partición recursiva que reemplaza las funciones de salto por una potencia truncada. Así se obtiene una aproximación en forma de expansión de

productos tensores de funciones base de tipo spline, el concepto de continuidad de la aproximación contenida en el producto tensorial está determinada por la elección del orden q para las funciones splines univariantes.

Una desventaja de la regresión spline es el llamado efecto extremo que consiste en que se producen largas contribuciones del error cuadrático medio, al estar x muy cerca de los bordes del dominio de D. Una solución es cambiar las funciones spline de tal modo que cerca de los extremos de los intervalos, estos se unan por una función lineal, que se asemeja a una función spline de orden $q = 1$ con derivadas continuas. En otros términos se reemplaza cada función $b\left(\frac{x}{s}, t\right) = (s(x - t))$ por funciones cúbicas truncadas

$$C(\tfrac{x}{S} = +1, t_-, t, t_+) = \begin{cases} 0 & , x \le t_- \\ p_+(x - t_-)^2 + r_+ (x - t_-)^3, & t_- < x < t_+ \\ (x - t) & , x \le t \end{cases} \quad , (19)$$

$$C(\tfrac{x}{S} = +1, t_-, t, t_+) = \begin{cases} -(x - t) & , x \le t_- \\ p_-(x - t_+)^2 + r_-(x - t_+)^3, & t_- < x < t_+ \\ 0 & , x \ge t \end{cases} \quad , (20)$$

Donde $t_- < t < t_+$, y

$$p_+ = \frac{(2t_+ + t_- - 3t)}{(t_+ - t_-)^2} \quad p_- = \frac{(3t_- - 2t_- - t_+)}{(t_- - t_+)^2} \quad , (21)$$

$$r_+ = \frac{(2t - t_+ - t_-)}{(t_+ - t_-)^3} \quad r_- = \frac{(t_- + t_+ - 2t)}{(t_- - t_+)^3} \quad , (22)$$

Así $C(\frac{x}{S} = +1, t_-, t, t_+)$ será continua y sus primeras derivadas también. Las funciones lineales truncadas $b\left(\frac{x}{s}, t\right)$ tendrán una sola ubicación del nodo t mientras que la función cúbica truncada tiene tres nodos: t (central) y $t_+ y\ t_-$(laterales)

MARS en cuanto a la colinealidad de los predictores, genera una serie de modelos que incrementan su orden de interacción y compara el valor de la validación cruzada generalizada.

El algoritmo MARS resulta de la selección de una sub-base del producto tensorial completo de las funciones base spline de las p variables, con nodos diferentes de los datos marginales, obteniendo una función base:

$$B_m(x) = \prod_{K=1}^{K_m} [\, S_{k_m}\left(x_{v(k,m)} - t_{k_m}\right)\,]^2 \qquad , (23)$$

Siendo K_m: n° de factores en la m-ésima función base.

S_{k_m}: toma sólo dos valores, +1 o -1, indicando el sentido izquierdo o derecho del truncamiento.

$v(k, m)$: etiqueta de la variable predictora 1 ≤ $v(k, m)$ ≤ p.

t_{k_m}: ubicación del nodo en cada una de las variables.

Exponente q: orden de la aproximación splines.

Los dos lados de la base potencia truncada equivalen a las ecuaciones del producto tensorial truncado, al incluir los monomios $\{x_j^K\}_{k=1}^{q=1\ldots p}{}_{j=1}$ en cada una de las variables y la constante $B_0(x) = 1$.

MARS selecciona un conjunto de funciones base usando la estrategia stepwise (forward/backward). Forward se refiere al proceso iterativo mediante el cual cada

interacción forma una lista expandida de funciones base las que se consideran primero, de manera simultánea y luego se escogen algunas.

Se van agregando al modelo dos funciones nuevas en cada iteración hasta tener un alto número de ellas, lo que se conoce como sobreajuste.

Forward parte con la función base $B_0(x) = 1$, luego de la M-ésima iteración se tiene $2M + 1$ funciones en el modelo $\{B_0(x) = 1\}_0^{2M}$.

$B_m(x)$ tiene la forma $B_m(x) = \prod_{K=1}^{K_m}[S_{km}(x_v\,(k,m) - t_{km})]^q$ y la $M + 1$ iteración agrega dos nuevas funciones base:

$$B_{2M+1}(x) = B_{l(M+1)}[+(x_{v(M+1)} - t_{M+1})]_+^q \quad , (24)$$

$$B_{2M+2}(x) = B_{v(M+1)}[-(x_{v(M+1)} - t_{M+1})]_+^q \quad , (25)$$

$B_{l(M+1)}(x)$ corresponde a una de las $2M + 1$ funciones base ya escogidas con $0 \leq l(M + 1) \leq 2M$.

$v(M + 1)$ es una de las variables predictoras.

t_{M+1} ubicación del nodo en la variable.

$l(M + 1), v(M + 1)$ y t_{M+1} son parámetros que generan las dos nuevas funciones base elegidas para un mejor ajuste del modelo, lo cual está definido por:

$$l(M+1), v(M+1), t_{M-1} = \arg min_{i=1}^{\sum N} \{\, y_i - \sum_{m=0}^{2M} a_m\, B_m(x) - a_{2M+1}\, B_l\,(x)\, [+(x_v - t)]^q - a_{2M+1}\, B_l\,(x)\ [-(x_v - t)]_+^q\}_{l,v,t}^2 \{a_m\}_0^{2M+2} \quad , (26)$$

En el algoritmo MARS ingresan las funciones base de menor orden de interacción antes de las de mayor orden.

La selección del modelo se realiza escogiendo un gran número de funciones base para luego ir eliminando las que están en exceso en una estrategia análoga a una regresión lineal estándar en la cual el proceso de elección backward se inicia con M_{max}funciones base que son el stock de variables factibles de ser elegidas luego de eliminar las funciones excesivas.

El modelo estima la falta de ajuste sobre un conjunto de datos representativos que no forman parte de la muestra de entrenamiento. Aquel modelo que logre minimizar el criterio de selección stepwise backward se toma como la estimación final de la función. Al ser MARS un procedimiento no lineal, se justifica un criterio basado en muestras reutilizables como validación cruzada (CV) o bootstrapping y que se define por: , (27)

$$CV = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}_{M/t}(x_i))^2$$

$\hat{f}_{M/t}$ corresponde a la M función base del modelo considerada en el procedimiento forward backward estimado con la i-ésima observación removida. Debido a la estructura jerárquica del grupo de modelos considerados en el proceso stepwise, lo anterior se evalúa para todos los $0 \leq M \leq M_{max}$ modelos.

El método de validación cruzada implica una replicación del modelo N veces con cada una de las observaciones removidas, sin embargo el criterio de validación cruzada generalizada (*GCV*) evalúa el modelo solo una vez:

$$GCV(M) = \frac{1}{N}\sum_{i}(y_i - \hat{f}_q\,(x_i))^2/[1 - \frac{C(M)}{N}]^2 \quad , (28)$$

$\hat{f}_q$: la dependencia de esta función y el criterio sobre el número de funciones base M es indicado explícitamente.

GCV: el numerador en esta expresión es error cuadrado medio del ajuste de los datos, mientras que el denominador corresponde a un término penalizado que representa el

incremento de la varianza asociada con el aumento de la complejidad del modelo (n° de funciones base M).

En el caso de que los parámetros de las funciones base del modelo MARS fueran definidos de manera independiente de los valores respuesta $(y_1, \ldots\ldots, y_N)$, entonces solo los coeficientes $(a_0, \ldots\ldots, a_M)$ están siendo ajustados por los datos. Así la complejidad de la función de costo está dada por:

$$C(M) = traza(B(B'B)^{-1}y_1B') + 1 \quad , (29)$$

B: matriz de datos MxN de las M funciones base, lo que es igual al número de funciones base linealmente independientes, por lo que $C(M)$ es el número de parámetros que está siendo ajustado.

Se sugiere utilizar la fórmula de $GCV(M)$ como criterio de validación, pero con un aumento de la función de complejidad de costo $C(M)$ para así reflejar los parámetros adicionales que según los coeficientes de expansión $(a_0, \ldots\ldots, a_M)$ están siendo ajustados para los datos. Así la función de complejidad de costo se expresa como:

$$C^{:}(M) = C(M) + d\ x\ M \quad , (30)$$

d es un costo para cada función base optimizada y es un parámetro del procedimiento.

M es el número de funciones base.

Al aplicarse el algoritmo MARS es posible obtener el siguiente modelo:

$$\hat{f}_q(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{K=1}^{K_m} [S_{k_m}(x_{v(k,m)} - t_{k_m})]_+ \quad , (31)$$

a_0: coeficiente de la función base constante B_1

La sumatoria incluye todas las funciones base B_m que se mantuvieron después de aplicar la estrategia backward y $S_{k_m} = \pm 1$. Con esta representación del modelo no se tiene una visión sobre la naturaleza de la aproximación, lo cual se soluciona reestructurando los

términos del modelo de manera que entregue información sobre la relación predictiva entre la variable respuesta y las variables predictoras o covariables. Así, el modelo puede formularse como sigue:

$$\hat{f}_q(x) = a_0 + \sum_{Km=1} f_i(x_i) + \sum_{Km=2} f_{ij}(x_i, x_j) + \sum_{Km=3} f_{ijk}(x_i, x_j, x_k) + \ldots..$$

La primera sumatoria corresponde a todas las funciones base con una sola variable, la segunda corresponde a la suma de las funciones base que incluyen dos variables, representando la interacción entre las dos variables y la tercera sumatoria es la contribución del efecto de interacción entre tres variables y así sucesivamente.

Con $V(m) = \{v(k,m)\}_1^{K_m}$ como el conjunto de variables asociadas con la m-ésima función base B_m, cada función de la primera sumatoria se expresa como:

$$f_i(x_i) = \sum_{\substack{K=1 \\ i \in V(m)}} a_m B_m(x_i) \qquad \text{, (33)}$$

Siendo ésta una suma sobre todas las funciones base que poseen una sola variable x_i y es una representación spline $q = 1$ de una función univariante. La función bivariante de la segunda sumatoria se expresa como:

$$f_{ij}(x_i, x_j) = \sum_{\substack{K=2 \\ (i,j) \in V(m)}} a_m B_m(x_i, x_j) \qquad \text{, (34)}$$

Siendo esta última la suma sobre todas las funciones base que incluyen dos variables x_i y x_j. Agregando esto a las correspondientes contribuciones univariantes, se obtiene:

$$f_{ij}^*(x_i, x_j) = f_i(x_i) + f_j(x_j) + f_{ij}(x_{ij}) \qquad \text{, (35)}$$

Siendo $q = 1$ la aproximación del producto tensor spline representando la contribución bivariada conjunta de x_i y x_j para el modelo. Términos que involucran más variables se representan de manera similar.

La descomposición de la tabla ANOVA facilita la interpretación del modelo MARS: identifica las variables que ingresan al modelo, si lo hacen aditivamente o involucran interacciones con otras variables y el nivel de las interacciones de las otras variables que participan.

Capítulo III: Objetivos general y específicos

3.1. Objetivo General

Proponer un modelo estadístico que permita encontrar la mejor discriminación del sexo en osamentas humanas a partir de variables métricas de clavícula.

3.2. Objetivos Específicos

Construir un modelo estadístico que permita predecir el sexo en osamentas humanas, en base a variables métricas de clavícula, aplicando las técnicas estadísticas Regresión Logística (Logit) y MARS.

Comparar ambas técnicas en cuanto a sus indicadores de bondad de ajuste con el objeto de detectar cuál de estas genera un mejor rendimiento.

Definir qué variables independientes (dimensiones de clavícula) son importantes en la determinación del sexo en osamentas.

Capítulo IV: Hipótesis de Investigación

La metodología MARS (Multivariate Adaptive Regression Spline) entregará mejores ajustes que la técnica de Regresión Logística, puesto que la primera herramienta utiliza funciones de suavizamiento que permite detectar nodos de corte que permitirán asociar mejor los predictores con la respuesta de estudio.

Capítulo V: Metodología

5.1 Diseño del estudio

Esta investigación corresponde a un estudio de tipo transversal retrospectivo, puesto que se analizarán las variables simultáneamente, en un momento acotado con datos de un tiempo pretérito [26]. Es además un diseño analítico, que busca descubrir la relación entre las variables continuas, largo máximo y circunferencia del eje de la clavícula y el sexo de los individuos analizados. Se trabajará con una cohorte de esqueletos humanos exhumados con fines de investigación.

5.2 Universo y muestra

El universo al cual pertenecen los restos óseos utilizados en esta investigación corresponde al grupo de fallecidos sepultados en los patios temporales del Cementerio General de Santiago.

Del Patio temporal N° 152 del Cementerio General (individuos fallecidos en los años 1997 y 1999) se constituyó una muestra a demanda (muestreo no probabilístico por conveniencia) de n=122 individuos esqueletizados que forman parte de la Colección de Docencia de la Unidad Especial de Identificación Forense (UEIF), del Servicio Médico Legal (SML) de Santiago. La muestra se compone de 46 individuos femeninos y 76 masculinos. Ciento cuatro (104) individuos poseen ambas clavículas y 18, solo una pieza (10 cuentan solo con la pieza izquierda y 8 con la derecha).

La muestra obtenida con un n=122, como ya se mencionó es a demanda, puesto que de un patio temporal de cerca de tres mil individuos se seleccionaron aquellos que tenían al menos una clavícula con alguna de las dimensiones a registrar, por lo que el número de individuos dependió exclusivamente de la disponibilidad y el estado de los restos, en otras palabras es una muestra que está constituida por todas las unidades de análisis que se pueden alcanzar para este estudio.

5.3 Variables y tópicos de estudio

La determinación del sexo a partir de dimensiones de clavícula es el tópico general de este análisis y las variables que se examinarán se muestran en la siguiente tabla (ver Tabla 2):

Tabla 2. Variables a considerar en el análisis.

Variable	Clasificación de la variable	Descripción
Sexo	Cualitativa, nominal, dicotómica, binaria, dependiente (respuesta)	Variable recodificada: Masculino=1 Femenino=0.
LongitudDer	Cuantitativa, de razón, continua, independiente (explicativa)	Corresponde a la longitud máxima de la clavícula derecha (desde sus extremos más prominentes) medida en milímetros.

LongitudIzq	Cuantitativa, de razón, continua, independiente (explicativa)	Corresponde a la longitud máxima de la clavícula izquierda (desde sus extremos más prominentes) medida en milímetros.
CircunDer	Cuantitativa, de razón, continua, independiente (explicativa)	Corresponde al contorno o circunferencia de la porción media del eje de la clavícula derecha medida en milímetros.
CircunIzq	Cuantitativa, de razón, continua, independiente (explicativa)	Corresponde al contorno o circunferencia de la porción media del eje de la clavícula izquierda medida en milímetros.

Cabe mencionar que las variables que corresponden a dimensiones de clavícula consideradas en este estudio, son las mismas que fueron analizadas en el artículo original del método a excepción de la variable peso de la pieza, que en esta investigación no se incluyó, puesto que podría estar severamente alterada por agentes tafonómicos del lugar de inhumación de los restos.

En cuanto a la edad de los individuos, esta variable se estudiará en términos de su relación con la variable respuesta y con las variables predictoras que se incluirán en los modelos, puntualmente se verá si corresponde a una variable confusora en la relación entre dimensiones de clavícula y sexo. Es preciso aclarar que se busca construir un modelo que permita discriminar el sexo en individuos no identificados con similares

características a la muestra que se seleccionó para este estudio y que por lo tanto no registran información de su edad. Sin embargo, es pertinente hacer un examen exploratorio previo para ver si alteraría la variable respuesta.

5.4 Actividades principales del trabajo de terreno

En el mes de Agosto de 2013, se efectuaron las gestiones pertinentes para la toma de muestras en el Cementerio General, a través de la Facultad de Medicina de la Universidad de Chile. Esta recolección de restos cadavéricos se realizó con el objeto de conformar una colección de material de estudio para este proyecto y futuros estudios.

5.5 Métodos o procedimientos de recolección de la información e instrumentos usados

Se confeccionó una ficha de registro de información para clavículas. Este documento se utilizó al momento de tomar cada muestra extraída desde el cementerio, registrando información del sitio de procedencia (patio y sepultura) y elemento óseo extraído. Posteriormente, un funcionario del cementerio, recopiló antecedentes del fallecido al cual pertenecía la muestra, como el sexo y la edad, características que se detallan en el pase de sepultación.

Se construyó una base de datos en una plantilla Excel 2013 que contiene las siguientes columnas:

-Caja: n° de caja donde se encuentra guardada la(s) clavícula(s) en el depósito de osamentas de la UEIF del SML de Santiago.

- n° de protocolo: corresponde al código de identificación del individuo y se asigna a los fallecidos al momento de su ingreso a la UEIF.

- clavículas: indica si el individuo tiene ambas piezas (derecha e izquierda) o solo una de éstas.

- sexo del fallecido

- edad: edad al morir del fallecido.

- LongituDer: longitud máxima en milímetros de la clavícula derecha.

- LongitudIzq: longitud máxima en milímetros de la clavícula izquierda.

- CircunDer: circunferencia o contorno del eje medio de la clavícula derecha en milímetros.

- CircunIzq: circunferencia o contorno del eje medio de la clavícula izquierda en milímetros.

Para medir el largo máximo de la clavícula, se utilizó una tabla osteométrica de la marca GPM y para el registro del contorno, una cinta métrica del kit de instrumentos antropométricos de la misma marca.

5.6 Técnicas de procesamiento de la información

La información recogida en la plantilla Excel se exportó al editor de datos del software estadístico Stata 13. Se efectuó el Análisis Descriptivo o Exploratorio de la muestra total (n=122), dividido en un análisis univariado y un análisis bivariado. Posteriormente se seleccionó de manera aleatoria una muestra de Entrenamiento, dejando las observaciones restantes como muestra de Validación (estas muestras también se sometieron a análisis exploratorio). Este procedimiento se generó seleccionando el 70 % de los individuos (n=85) de la muestra original, dejando el resto de las unidades de observación, es decir el 30 % de los individuos (n = 37), como parte de la segunda muestra.

Con el mismo programa estadístico Stata 13 se aplicó Análisis de Regresión Logística a la muestra de Entrenamiento y para conocer la capacidad de discriminación del modelo obtenido se generaron los siguientes gráficos: curva ROC (*Receiver Operating Characteristic*) que expone la sensibilidad frente a la especificidad para un sistema

clasificador binario según se varía el umbral de discriminación; curvas de sensibilidad y especificidad y gráfico de caja y bigotes de los individuos masculinos y femeninos discriminados por la regresión. Además, se tabularon puntos de corte (probabilidad) junto a la sensibilidad, especificidad y casos correctamente clasificados con el objeto de hallar un punto de corte óptimo de discriminación, el cual se consideró en el modelo logístico para la variable respuesta (sexo). La información obtenida en este último punto fue también presentada en una tabla.

Posteriormente, y con el objetivo de legitimar el modelo, este fue aplicado a la muestra denominada Validación, tabulándose los indicadores de sensibilidad y especificidad correspondientes.

Se aplicó análisis MARS con el software del mismo nombre en su versión 8.0 desarrollado por Salford Systems a la muestra de Entrenamiento. Se generaron dos modelos: uno sin interacción y otro con dos interacciones cuyos resultados fueron expuestos en sus tablas correspondientes, detallando aspectos relevantes como funciones base, R^2 ajustado y GCV. Además, se expone la importancia que tiene cada variable en los modelos generados, exponiéndose esta jerarquía en sus tablas correspondientes.

En cuanto a los indicadores de bondad de ajuste, se aplicaron los dos modelos obtenidos a la muestra Validación y se construyeron tablas que exhiben valores de R^2 ajustado, GCV, sensibilidad, especificidad y casos correctamente clasificados para cada uno de los modelos obtenidos por análisis MARS.

Se presentó la tabla ANOVA del modelo escogido que expone las funciones del modelo y su respectiva desviación estándar, costo de omisión y variables involucradas. Adicionalmente, se entregó una descripción de cada una de las funciones base del modelo seleccionado, se expuso la ecuación del modelo como una combinación lineal de funciones base, junto al modelo explícito de éste. Se presentó el efecto que tienen las

variables predictoras sobre la variable respuesta presentando su contribución en los gráficos respectivos.

Finalmente, se efectuó la comparación del modelo generado mediante Regresión Logística con el modelo obtenido del análisis MARS, en cuanto a sus indicadores de bondad de ajuste, los cuales se expusieron en la tabla respectiva.

5.7 Aspectos vinculados a la obtención de la muestra.

Como ya se mencionó en el capítulo IV “Metodología”, en el año 2013, un grupo de peritos de la UEIF del SML dirigidos por la Antropóloga Física Paulina Marambio V. inició las gestiones pertinentes para crear una colección de osamentas humanas con fines de investigación, tendientes a la evaluación y validación de métodos de identificación en restos óseos que son usados habitualmente en el trabajo pericial. Dicha colección se conformó para este proyecto y para futuros trabajos de validación con el apoyo del Instituto Dr. Carlos Ybar (ente dependiente del SML), del Cementerio General y la colaboración del Programa de Anatomía y Biología del Desarrollo del Instituto de Ciencias Biomédicas (ICBM) de la Facultad de Medicina de la Universidad de Chile.

Capítulo VI: Resultados

6.1. Análisis Descriptivo o Caracterización de la Muestra

La muestra se compone de un total de n=122 unidades de análisis que corresponden a individuos adultos de ambos sexos, de los cuales se midieron las clavículas derecha e izquierda, siendo estas dimensiones las variables que se considerarán como independientes en la construcción del modelo.

6.1.1. Presencia de datos faltantes o *missing values*

Es importante mencionar la presencia de datos *missing* en las variables independientes del modelo. Se registraron las mediciones correspondientes a cada dimensión de clavícula para las piezas derecha e izquierda, considerando para este propósito, la muestra total (n=122). Para la variable Longitud máxima clavícula derecha, se obtuvieron 112 mediciones, equivalente al 91.8% de los individuos de la muestra. En cuanto a la variable Longitud máxima clavícula izquierda, las mediciones efectivas fueron 100, correspondiendo al 81.97% del total de individuos. Para Circunferencia clavícula derecha, se obtuvieron 116 mediciones (95.08% del total de individuos) y para la variable Circunferencia clavícula izquierda, 108 mediciones obtenidas, equivalentes al 88.52% de los individuos de la muestra. La frecuencia de datos faltantes en cada variable según sexo, se detalla en la Tabla 3.

Tabla 3. Frecuencia de datos *missing* en la muestra.

	Sexo				
	Masculino		Femenino		
Variable	n	%	N	%	Total
Longitud Máxima Clavícula Derecha	5	50	5	50	10
Longitud Máxima Clavícula Izquierda	12	54.5	10	45.5	22
Circunferencia Clavícula Derecha	3	50	3	50	6
Circunferencia Clavícula Izquierda	9	64.3	5	35.7	14

En las dos dimensiones examinadas, longitud y circunferencia, la frecuencia de datos faltantes en la pieza izquierda duplica a los del lado derecho. Considerando el total de estos datos en cada variable, se observa que un 69.03% de datos faltantes se asocia a la pieza izquierda.

6.1.2. Análisis Univariado

Sexo de los individuos

El sexo documentado de los individuos que componen la muestra se encuentra distribuido de la siguiente forma (Tabla 4):

Tabla 4. Distribución de la muestra por sexo.

Sexo	n	%
Masculino	76	62.3
Femenino	46	37.7
Total	122	100

Los individuos masculinos superan en un 25% a aquellos femeninos.

Edad de los individuos de la muestra

Las edades de los individuos de la muestra fluctúan entre los 16 y 96 años. Tanto el grupo de hombres como el de mujeres se conforman mayoritariamente por individuos de edades sobre los 49 años, tal como se exhibe en la Tabla 5 y en los gráficos de cajas y bigotes e histogramas (Figuras 1 y 2).

Tabla 5. Distribución de la edad de los individuos de la muestra por rango etario.

Rango de Edad (años)	n	%
10-19	1	0.8
20-29	3	2.61
30-39	8	7
40-49	14	12.17
50-59	13	11.3
60-69	19	16.52

70-79	30	26.1
80-89	20	17.4
90-99	7	6.1
Total	115	100

En la tabla se observa gran representación de individuos cuyas edades fluctúan entre los 59 y 89 años. Por otra parte, se aprecia baja representación de individuos jóvenes y muy seniles en la muestra analizada.

Tabla 6. Distribución de la edad por sexo.

	Sexo											
	Masculino						**Femenino**					
	n	**Mín-Máx**	**D.S.**	**Media**	**Mediana**	**R.I.**[2]	**n**	**Mín-Máx**	**D.S.**	**Media**	**Mediana**	**R.I.**
Edad	74	16 - 92	18.36	62.92	66.5	26	41	31 - 96	17.64	70.29	74	25

En la Tabla 6 se exponen medidas de tendencia central y medidas de dispersión de la variable edad, según sexo, además del número de individuos de cada grupo. Se observa que la cantidad de individuos masculinos supera considerablemente a los de sexo femenino. La media y la mediana de la variable edad para los individuos femeninos son superiores que para su contraparte masculina, mientras que la desviación estándar y el rango Intercuartílico muestran cifras similares.

[2] R.I. corresponde al Rango Intercuartílico.

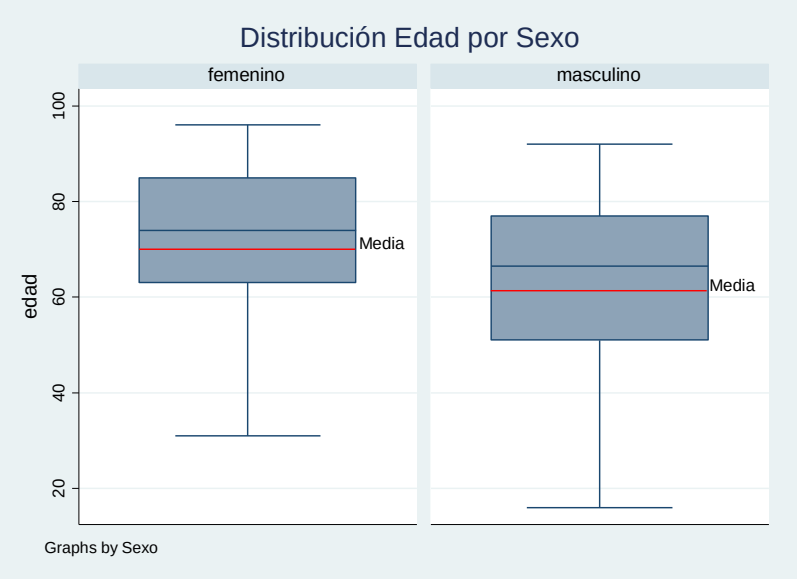


Figura 1. Gráficos de caja y bigotes con distribución de la edad por sexo.

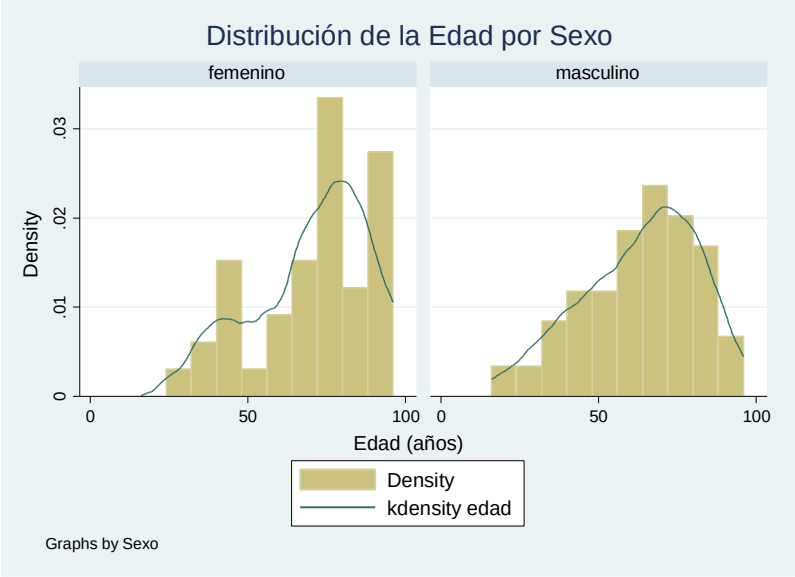


Figura 2. Histogramas que muestran la distribución de la edad de los individuos por sexo.

Para ambos sexos existe asimetría negativa y una gran dispersión en la distribución de los datos.

La Tabla 7 muestra los resultados del test de asimetría y curtosis y del test de normalidad de Shapiro-Wilk para la variable edad.

Tabla 7. Análisis de normalidad de la variable edad.

	Asimetría/Curtosis				Normalidad	
Variable	**Asimetría**	**Curtosis**	**chi2 (2)**	**Prob>chi2**	**W**	**p**
Edad	0.0124	0.2589	6.96	0.0308	0.95581	0.0008

En la tabla anterior se observa que el test de asimetría y curtosis rechaza la hipótesis nula de normalidad para la variable edad. En cuanto al test de normalidad, se registra un alto valor para W (cercano a 1), pero un valor $p<0.05$, por lo tanto se rechaza la hipótesis nula de que la variable edad distribuye normal.

Dimensiones de la Clavícula (variables independientes y predictoras en el modelo).

A continuación se exponen estadísticas descriptivas de las variables dimensiones de clavícula (Tabla 8).

Tabla 8. Estadísticas descriptivas de variables dimensiones de clavícula en muestra total.

Variable	n	%	Min - Máx	Media	D. S.	Mediana	R.I.
Longitud Máxima Clavícula Derecha	112	91.8	123.7 - 170.1	146.48	9.86	147.8	12.75
Longitud Máxima Clavícula Izquierda	100	81.97	127.3 - 171.5	148.04	10.04	147.9	14.87
Circunferencia Clavícula Derecha	116	95.09	22 - 49	36.47	4.87	36	7
Circunferencia Clavícula Izquierda	108	88.53	24 - 47	35.04	4.25	35	6

De acuerdo a la información contenida en la Tabla 8 se observa que para la dimensión longitud, los valores de la media, la mediana, la desviación estándar y el rango Intercuartílico son similares para la pieza derecha e izquierda. En cuanto a la dimensión circunferencia de clavícula, también se registran similitudes para estas medidas descriptivas entre piezas derecha e izquierda.

La distribución de cada una de las variables se observa en los siguientes gráficos (Figuras 3 y 4).

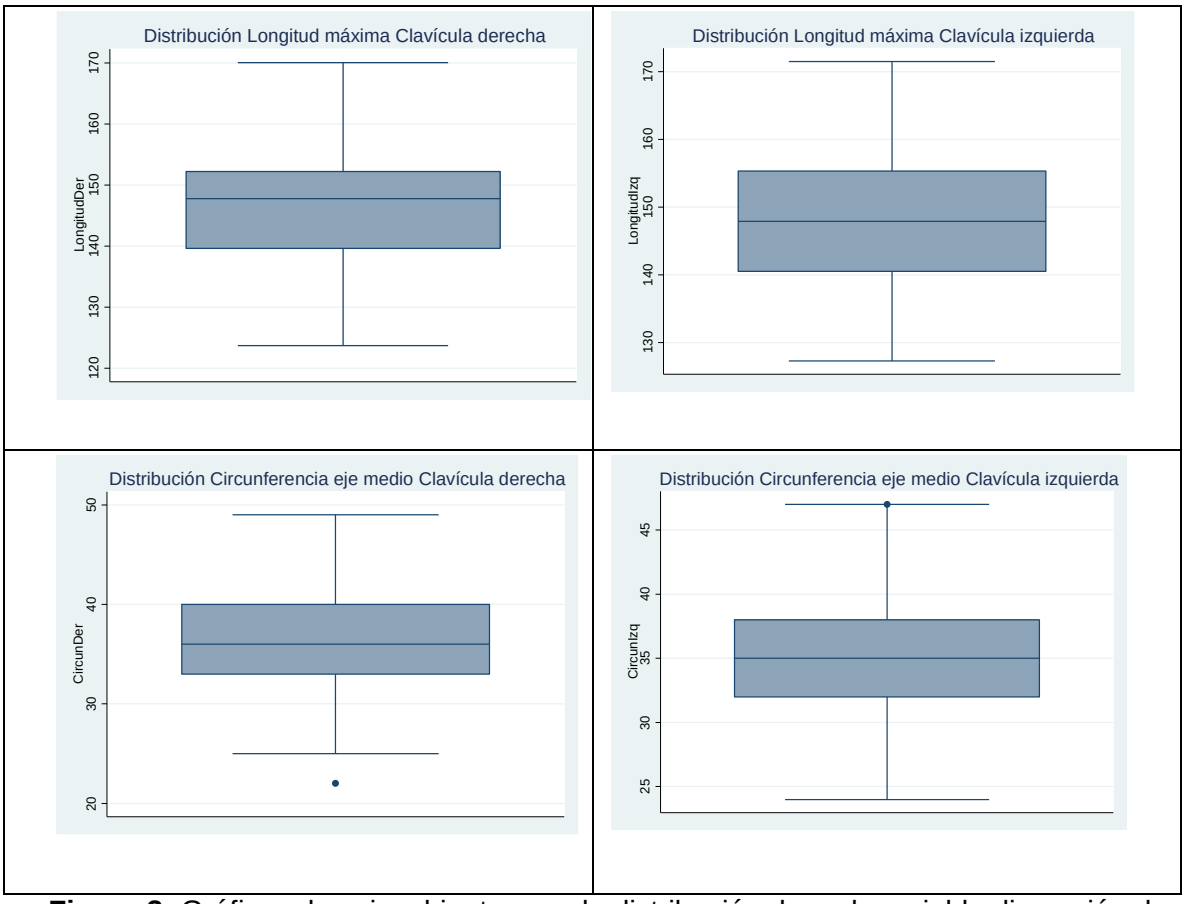
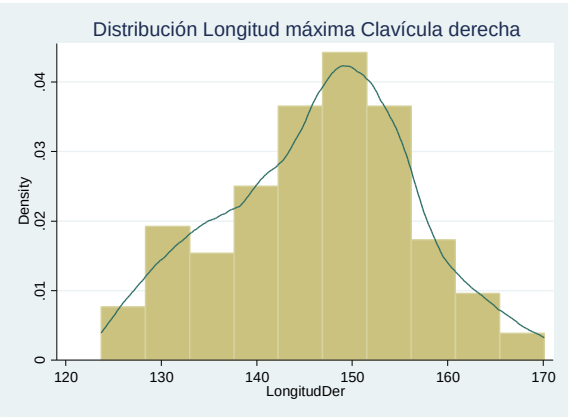
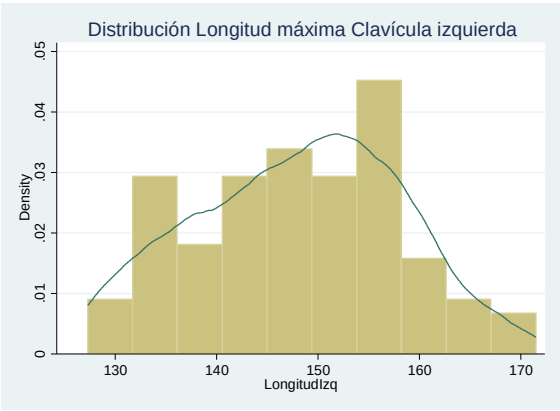


Figura 3: Gráficos de caja y bigotes con la distribución de cada variable dimensión de clavícula.

En la Figura 3, el gráfico de caja y bigotes de la variable Longitud máxima Clavícula derecha (gráfico superior izquierdo) muestra que la mediana (línea horizontal de la caja) no se sitúa al centro de la caja, lo cual indica que esta medida no coincide con la media (consistente con lo observado en la Tabla 8 para esta variable). Por otra parte, el 50 % de las observaciones de la variable se concentran en el rango de los 140 y 152 mm aproximadamente (valores correspondientes al borde inferior y superior de la caja, respectivamente). Otro 50% lo constituyen las longitudes bajo el valor correspondiente a la mediana. El 25% de la distribución de los valores de la variable se hallan bajo los 140 mm, mientras que otro 25% de los casos se encuentran sobre los 152 mm hasta mayor valor para la variable (170 mm). El gráfico de caja y bigotes de la variable Longitud

máxima Clavícula izquierda (gráfico superior derecho) muestra una simetría en la distribución de los valores de la variable (bigotes de similar longitud y mediana situada en el centro de la caja). Los extremos inferior y superior de los bigotes indican los valores mínimo y máximo de la variable, respectivamente. El borde inferior de la caja corresponde al primer cuartil, la línea horizontal dentro de la caja representa a la mediana y el borde superior de la caja corresponde al tercer cuartil de la distribución. El gráfico de caja y bigotes correspondiente a la variable Circunferencia eje medio de Clavícula derecha muestra simetría en la distribución de la variable, sin embargo se observa un dato atípico ubicado bajo la distribución de la variable. Para la variable Circunferencia eje medio Clavícula izquierda, el gráfico respectivo muestra que la distribución de la variable es simétrica y que el valor de la mediana coincide con el valor de la media.





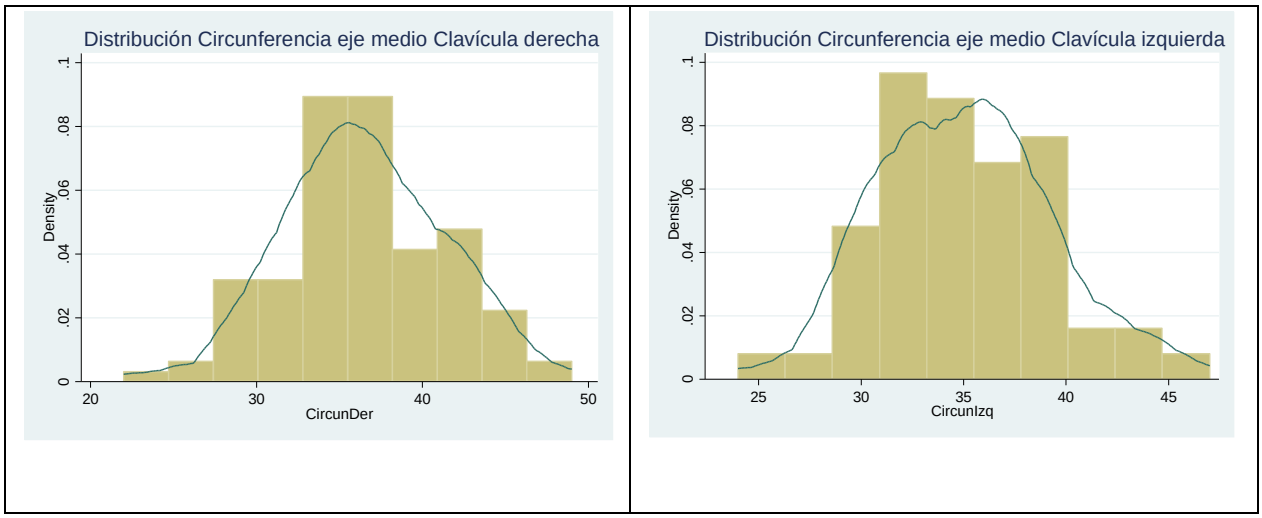


Figura 4. Distribución de cada una de las variables que serán independientes en los modelos.

La Figura 4 presenta histogramas para cada dimensión de clavícula. En general, se observa asimetría en la distribución de cada variable.

En la siguiente tabla (Tabla 9) se exponen más detalles respecto a la distribución de las variables dimensiones de clavícula

Tabla 9. Resultados de los test de asimetría, curtosis y normalidad de las variables dimensiones de clavícula.

	Asimetría/Curtosis				Normalidad	
Variable	**Asimetría**	**Curtosis**	**chi2 (2)**	**Prob>chi2**	**W**	**p**
Longitud máxima clavícula derecha	0.5533	0.4022	1.07	0.5847	0.98576	0.28304
Longitud máxima clavícula izquierda	0.6997	0.0588	3.83	0.1477	0.98224	0.19784
Circunferencia clavícula derecha	0.9822	0.7038	0.15	0.93	0.99627	0.99069
Circunferencia clavícula izquierda	0.2512	0.7327	1.46	0.4809	0.99196	0.77952

Se observa para todas las variables expuestas, que tanto el test de asimetría y curtosis como el test de normalidad de Shapiro-Wilk, no rechaza la hipótesis nula de normalidad para la distribución.

La Tabla 10 muestra estadísticas de cada una de las variables predictoras según el sexo.

Tabla 10. Estadísticas descriptivas de las variables predictoras según sexo.

Variable	Sexo													
	Masculino							Femenino						
	n	%	Min - Máx	Media	D. S.	Mediana	R.I.	n	%	Mín-Máx	Media	D. S.	Mediana	R.I.
Longitud Máxima Clavícula Derecha	71	93.42	134.4-170.1	151.28	6.94	151.3	8	41	89.13	123.7-167.7	138.16	8.59	137.7	11.95
Longitud Máxima Clavícula Izquierda	64	84.21	134.2-171.5	153.22	7.3	154	9.5	36	78.26	127.3-160.7	138.83	7.29	138.55	10.9
Circunferencia Clavícula Derecha	73	96.05	30-49	38.74	3.94	38	6	43	93.48	22-40	32.63	3.76	33	5
Circunferencia Clavícula Izquierda	67	88.16	30-47	37.1	3.56	37	4	41	89.13	24-38	31.66	2.91	31	4

Los resultados de la Tabla 10 indican que las dimensiones examinadas (longitud y circunferencia) de la clavícula tanto derecha como izquierda, son mayores en los individuos masculinos. Para este mismo sexo, las variables Longitud Máxima Clavícula Derecha y Longitud Máxima Clavícula Izquierda, muestran valores mínimos y máximos muy similares, sin embargo, los valores de las otras medidas descriptivos muestran diferencias notorias. En cuanto al sexo femenino, y tanto para la dimensión longitud como para circunferencia, existen diferencias en los valores de medidas descriptivas expuestas, entre piezas derechas e izquierdas, a excepción de las medias registradas en las variables longitud, que prácticamente muestra el mismo valor.

A continuación se presenta en gráficos de caja y bigotes y en histogramas, la desagregación de cada variable por sexo.

Para Longitud máxima Clavícula derecha (ver Figuras 5 y 6) se tiene la siguiente distribución de la variable por sexo.

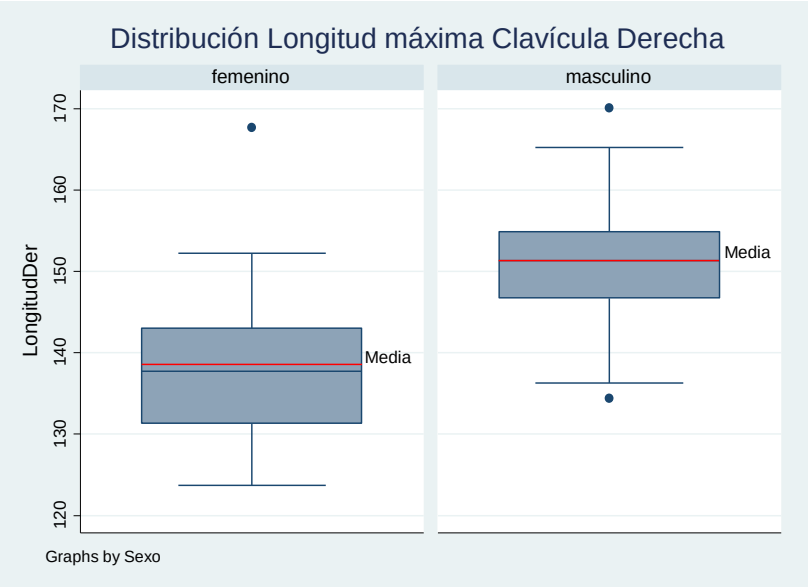


Figura 5. Gráficos de caja y bigotes de la variable por sexo.

Ambos sexos muestran una distribución que tiende a la simetría, pero con la existencia de datos atípicos u observaciones outliers.

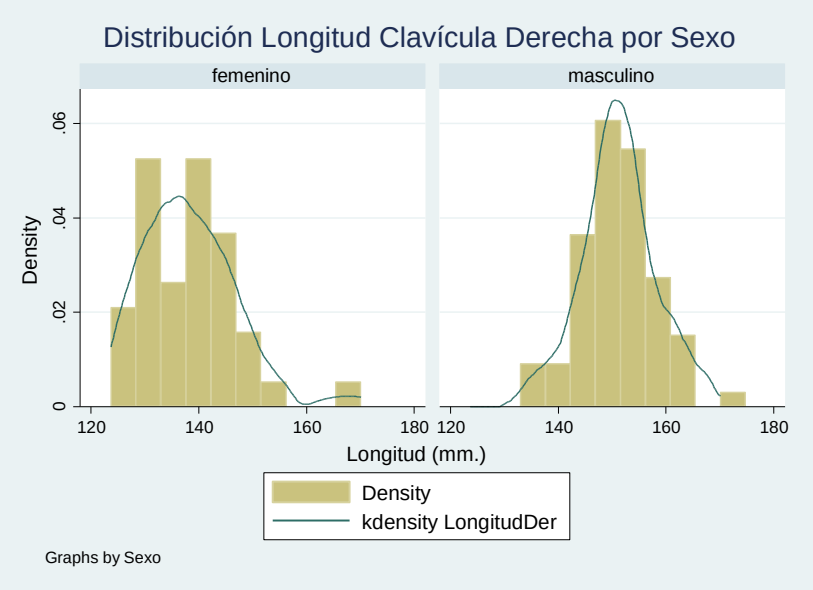


Figura 6. Distribución de la variable por sexo.

El histograma que representa la distribución de la variable examinada en el sexo femenino exhibe asimetría positiva, mientras que para el sexo masculino se observa una curva leptocúrtica.

Para la variable Longitud máxima Clavícula izquierda la distribución por sexo es la siguiente (Figura 7 y 8).

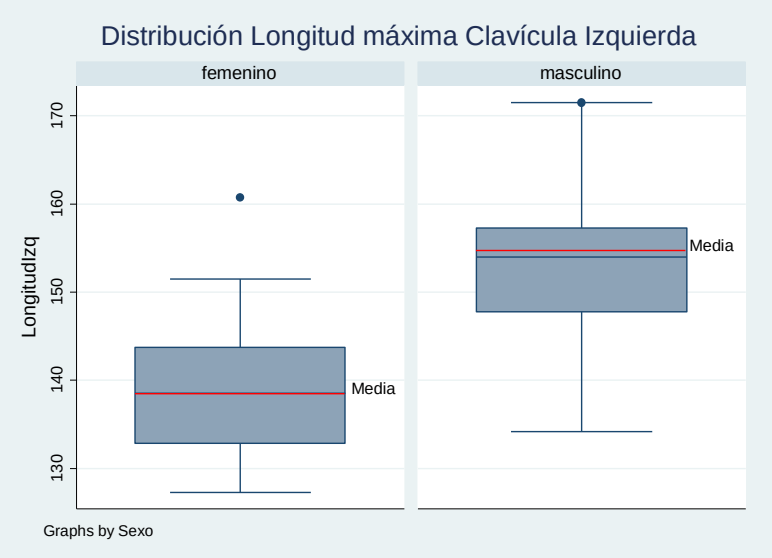


Figura 7. Gráficos de caja y bigotes de la variable por sexo.

Se observa para los individuos de sexo femenino, menores dimensiones en la variable estudiada, que aquellas registradas en el grupo masculino. La distribución de las longitudes en el sexo femenino se caracteriza por ser muy simétrica, sin embargo se observa la existencia de un dato atípico (muy superior al valor máximo de la variable).

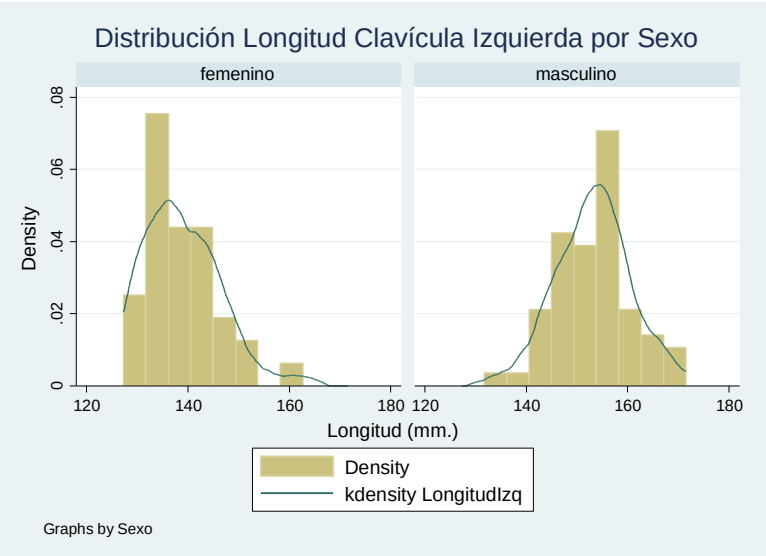


Figura 8. Distribución de la variable por sexo.

Los histogramas muestran asimetría en la distribución de los datos para ambos sexos. En el sexo femenino se observa una gran concentración de individuos con longitudes de clavícula izquierda cercanas a los 131 mm, mientras que en el sexo masculino existe alta concentración de longitudes próximas a los 159 mm.

Para la variable Circunferencia del eje medio de clavícula derecha la distribución es la siguiente (Figuras 9 y 10).

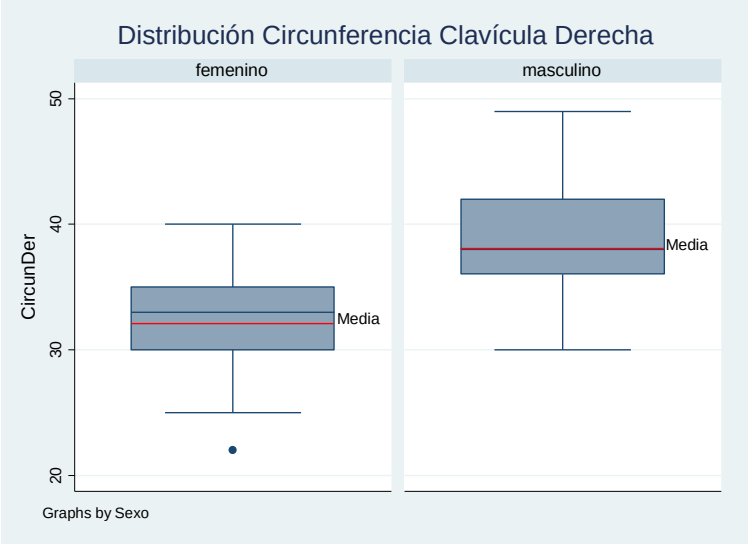


Figura 9. Gráficos de caja y bigotes de la variable por sexo.

En individuos femeninos se observa una distribución que tiende a la simetría, a pesar de la existencia de datos outliers y una media y mediana que no coinciden en sus valores. El gráfico de caja y bigotes para los individuos masculinos muestra en general una distribución que también tiende a la simetría, sin embargo se observa que la caja del diagrama muestra menor dispersión de los datos entre el primer y segundo cuartil de la distribución.

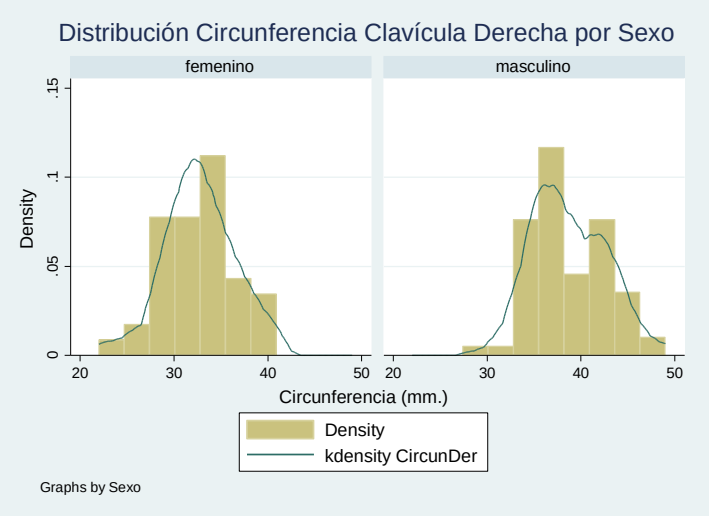


Figura 10. Distribución de la variable por sexo.

Igualmente los histogramas muestran una tendencia a la simetría de la curva para ambos sexos. En mujeres los individuos con circunferencias de eje de 35 mm son los más frecuentes y en el caso de los individuos masculinos las clavículas con circunferencias entre los 36 y 37 mm comprenden la mayoría.

Por último, para la variable Circunferencia del eje medio de la clavícula izquierda, la distribución es la siguiente (Figuras 11 y 12).

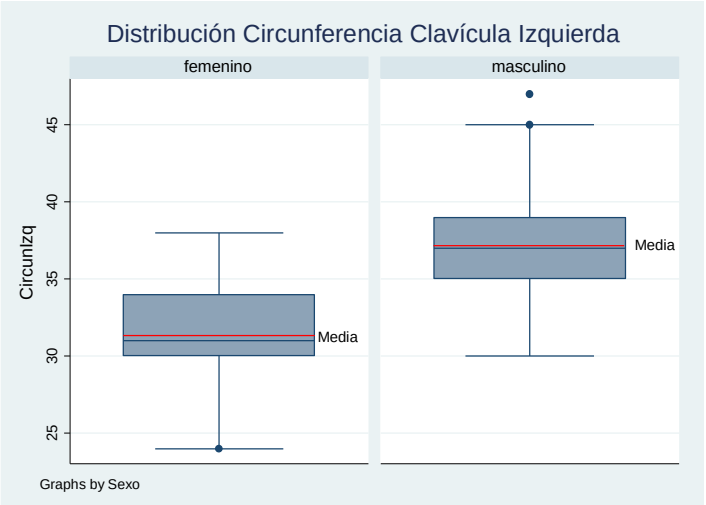


Figura 11. Gráficos de caja y bigotes de la variable según sexo.

Se observa en los gráficos de caja y bigotes que tanto para individuos masculinos como femeninos existe para la distribución de la variable expuesta, una tendencia a la simetría. Además, se aprecia que para el caso de los individuos femeninos, existe mayor dispersión de los datos entre el 2° y 3er cuartil (50% y 75%) de la distribución que entre el 25% y 50% de ésta.

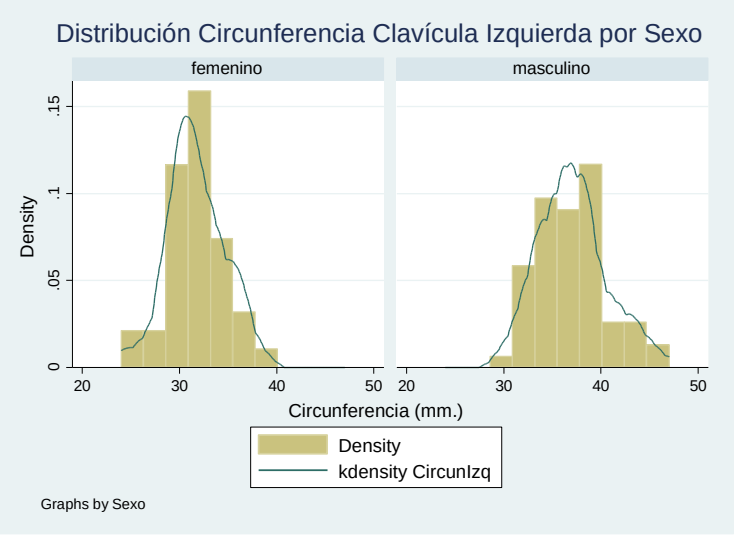


Figura 12. Distribución de la variable según sexo.

El histograma que representa la distribución de los individuos femeninos muestra que la mayor parte de las circunferencias se concentran entre los 32 y 33 mm, mientras que para individuos masculinos la mayor parte de los datos se concentra entre los 33 y 40 mm.

6.1.3. Análisis Bivariado

Para definir las variables a incorporar en el modelo se estudiará la relación entre el sexo, (variable respuesta del modelo) con dimensiones de clavícula (predictores), además de la relación del sexo y la variable edad.

6.1.3.1 Sexo y Dimensiones de Clavícula

Considerando los resultados de las pruebas de normalidad aplicadas a las variables dimensiones de clavícula, en las cuales no se rechazó la hipótesis nula de normalidad, se aplicó la prueba t-Student con el objeto de conocer si en cada variable independiente existe diferencia de medias entre individuos masculinos y femeninos. Los resultados de esta prueba se muestran en la siguiente tabla (Tabla 11):

Tabla 11. Resultados de la prueba t-Student en variables dimensiones de clavícula.

	Sexo						t-Student	
	Masculino			Femenino				
Variables	n	$\bar{x}$	S	n	$\bar{x}$	S	t	
Longitud Clavícula Derecha	71	151.28	6.94	41	138.15	8.59	-8.83	p<0.05
Longitud Clavícula Izquierda	64	153.22	7.3	36	138.83	7.29	-9.47	
Circunferencia Clavícula Derecha	73	38.74	3.94	43	32.63	3.76	-8.2	
Circunferencia Clavícula Izquierda	67	37.1	3.55	41	31.66	2.91	-8.25	

Se observó que para las cuatro variables evaluadas se rechazó la hipótesis nula de igualdad de medias entre los sexos femenino y masculino. En otras palabras se puede concluir que hay dependencia entre el sexo y cada una de las dimensiones de clavícula, siendo éstas buenos predictores del sexo.

6.1.3.2 Sexo y Edad

En el apartado 6.1.2. Análisis Univariado, el test de Shapiro-Wilk rechazó la hipótesis de normalidad para la variable edad (p=0.00080), sin embargo para aplicar el test t se debe cumplir la condición de normalidad o bien que ambos grupos a comparar (femenino y masculino) tengan más de 30 observaciones. Lo último se cumple para esta variable (n=41 para femenino y n=74 para masculino), por lo cual es posible aplicar el test t, previa

evaluación de la igualdad de varianzas. Los resultados de los test se muestran en la Tabla 12.

Tabla 12. Resultados de los test de varianzas y test t en variable edad.

Variable	p (test de varianza)	P (test t)
Edad	0.7981	0.0387

El test de varianzas señala que entre individuos femeninos y masculinos las varianzas no son distintas ($p > 0.05$) y el test t indica que las medias son iguales ($p < 0.05$).

En el caso de considerar el resultado de Shapiro-Wilk, que señala que la variable edad no es normal, sería pertinente aplicar el test de medianas cuyo resultado para p es 0.216, lo que indica que no hay una diferencia significativa entre las medianas de los grupos femenino y masculino que en otras palabras quiere decir que existe no independencia entre el sexo y la edad.

6.1.3.3 Estudio de la Correlación entre variables

Se procederá a estudiar mediante el coeficiente de correlación de Pearson qué variables predictoras están relacionadas, se considera como significativo un nivel del 5%.

La figura 13 muestra la relación entre las variables Longitud máxima de clavícula derecha y Longitud máxima de clavícula izquierda (Coeficiente de correlación, $r=0.9193$; Nivel de significación, $p<0.05$), existiendo una muy fuerte correlación directa.

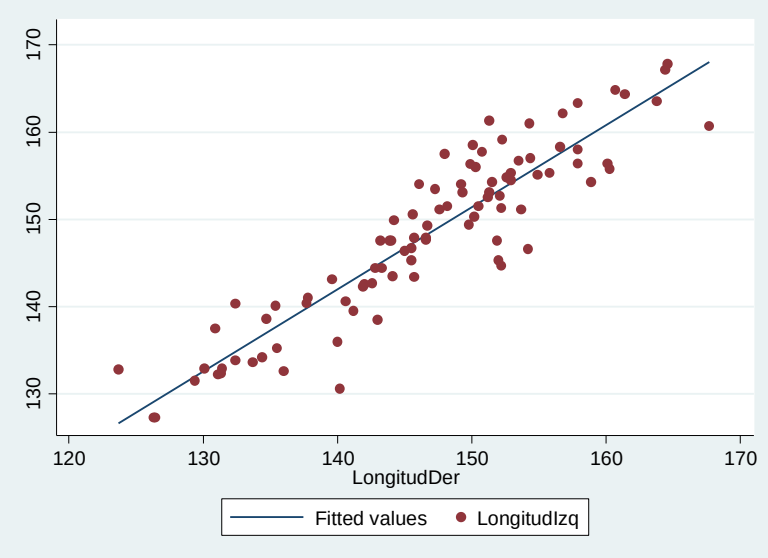


Figura 13. Diagrama de dispersión variables Longitud derecha y Longitud izquierda.

Para las variables Longitud máxima clavícula derecha y Circunferencia eje medio de clavícula derecha, los valores de r y p (r=0.4262; p<0.05) y el diagrama de dispersión de la figura 14, muestran una correlación moderada y positiva.

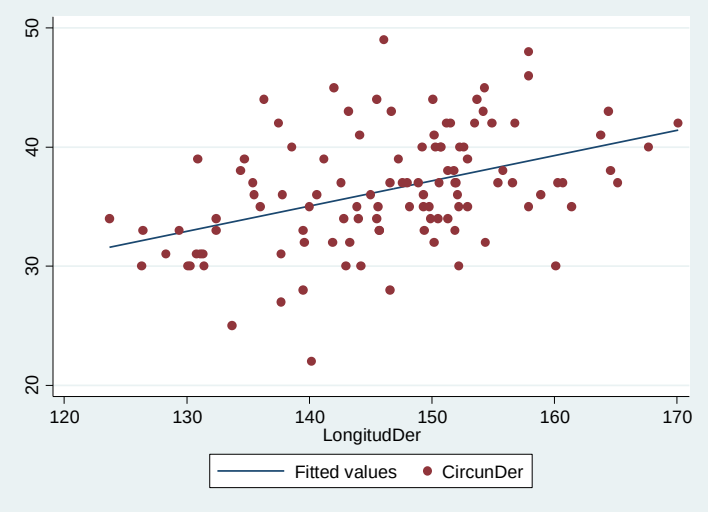


Figura 14. Diagrama de dispersión que muestra correlación entre Longitud derecha y Circunferencia derecha.

Los valores de r y p (r=0.4598; p<0.05) y la Figura 15, muestran que entre las variables Longitud derecha y Circunferencia izquierda, existe correlación moderada y directa.

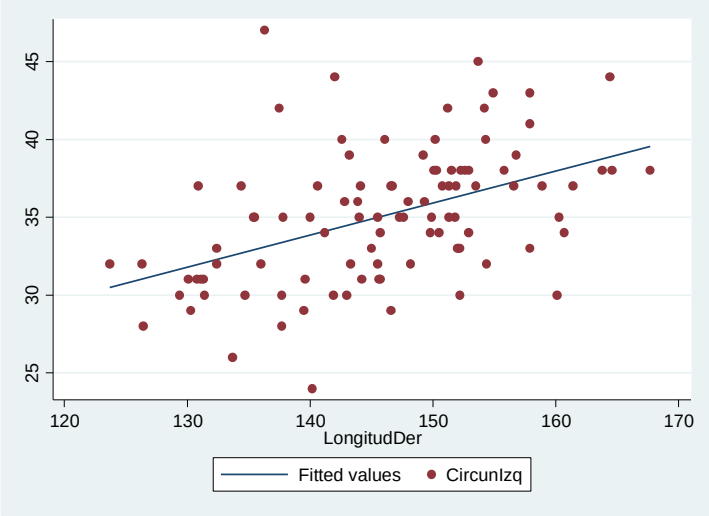


Figura 15. Gráfico de dispersión exhibiendo correlación entre Longitud derecha y Circunferencia izquierda.

La figura 16 junto con los valores de r=0.5315 y p<0.05, muestran una correlación moderada y directa entre las variables Longitud máxima de clavícula izquierda y Circunferencia del eje de clavícula derecha.

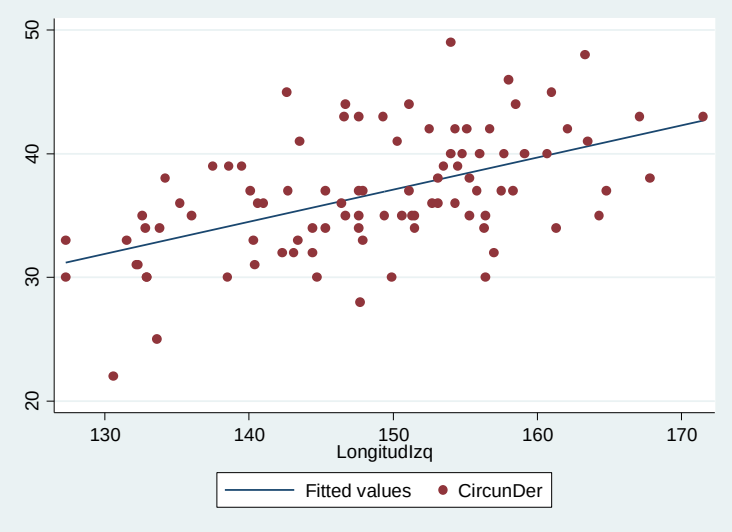


Figura 16. Diagrama de dispersión de correlación entre las variables Longitud Izquierda y Circunferencia derecha.

La figura 17 exhibe una correlación positiva y moderada (r=0.5651; p<0.05) entre las variables Longitud máxima izquierda y Circunferencia del eje izquierda.

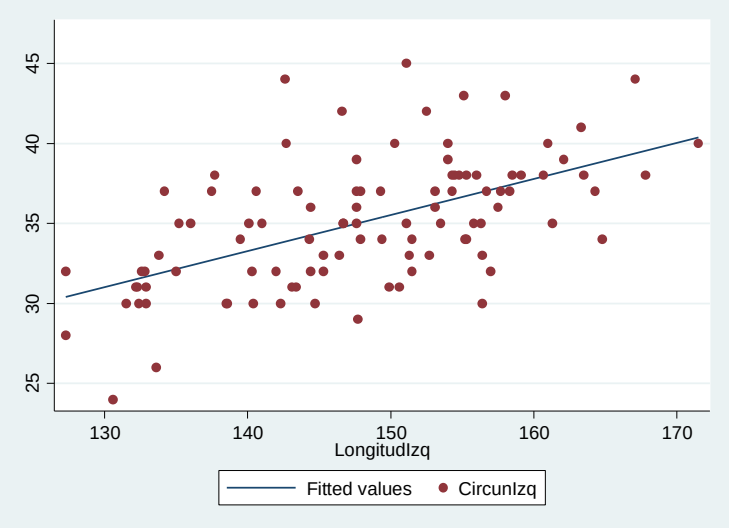


Figura 17. Gráfico de dispersión de la correlación entre variables Longitud izquierda y Circunferencia izquierda.

Finalmente la Figura 18 junto con los valores de r=0.8774 y p<0.05, indican una correlación muy fuerte y positiva entre las variables Circunferencia del eje de clavícula derecha y Circunferencia del eje de clavícula izquierda.

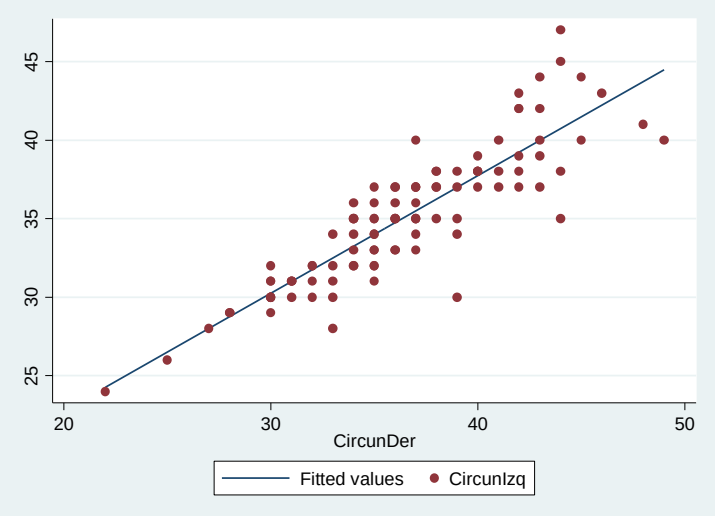


Figura 18. Gráfico de dispersión que exhibe la correlación entre las variables Circunferencia derecha y Circunferencia izquierda.

Es importante destacar la fuerte correlación que existe entre la misma dimensión de clavícula pero de lados distintos, esto es: entre Longitud derecha con Longitud izquierda y entre Circunferencia derecha con Circunferencia izquierda.

Se estudió la correlación entre la variable edad y las variables dimensiones de clavícula y se halló que no existe correlación entre esta variable y las cuatro variables predictoras. Los resultados se exponen en la Tabla 13 y la Figura 19.

Tabla 13. Valores en la correlación entre la variable edad con dimensiones de clavícula.

	Longitud Derecha		Longitud Izquierda		Circunferencia Derecha		Circunferencia Izquierda	
	R	**P**	**r**	**p**	**r**	**p**	**r**	**p**
Edad	-0.1856	0.0568	-0.1169	0.2646	-0.0106	0.9124	0.118	0.2401

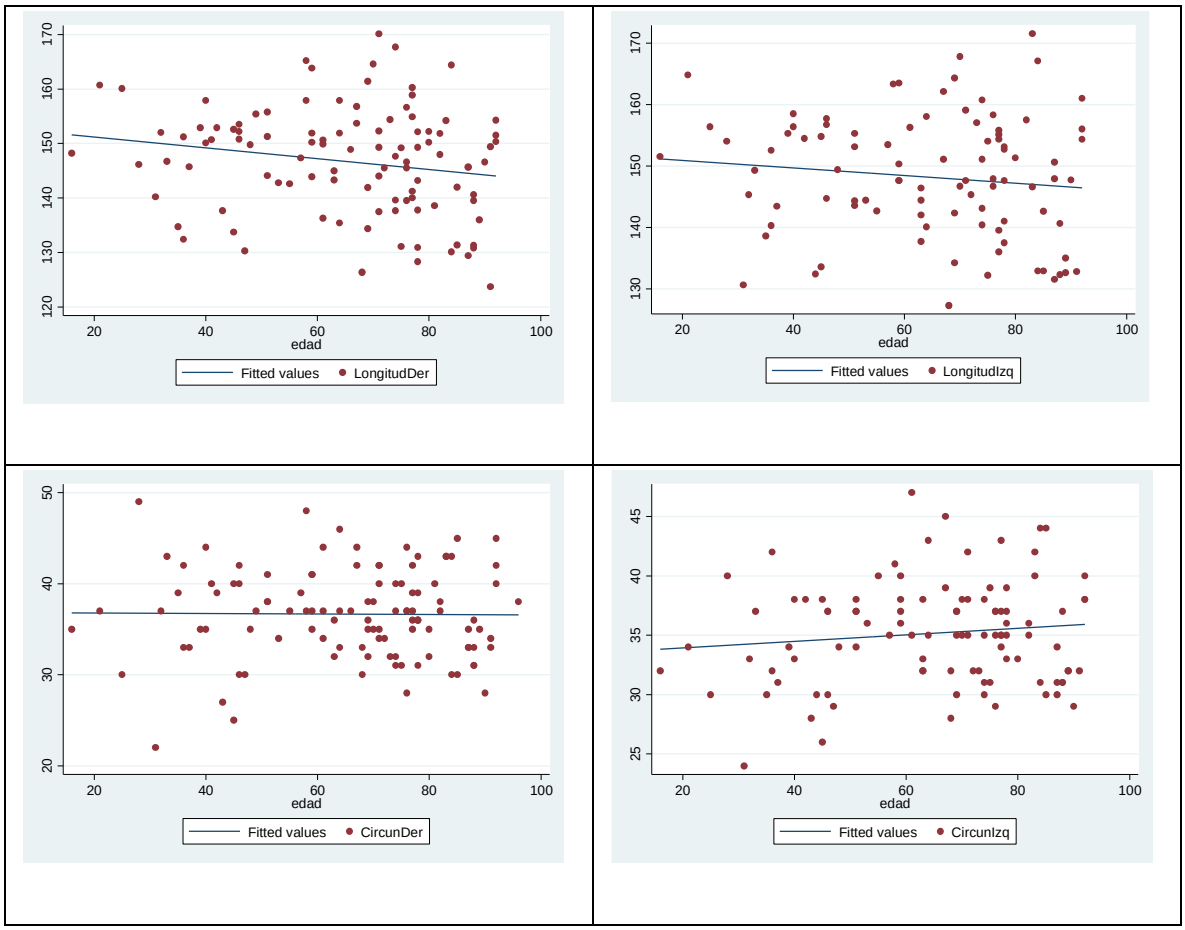


Figura 19. Gráficos de dispersión que muestran correlación inexistente entre las variables edad y dimensiones de clavícula.

Los resultados del análisis bivariado señalan que existe dependencia entre la variable sexo y las variables dimensiones de clavícula. Sin embargo, esta relación no existe entre

la variable sexo y edad. Por otra parte, la no existencia de correlación entre la variable edad y las variables dimensiones de clavícula descarta a la edad como variable confusora.

De lo anterior se concluye que solo las variables sexo y dimensiones de clavícula serán incorporados en el modelo.

6.1.4. Análisis Descriptivo Muestra Entrenamiento.

La muestra de Entrenamiento es aquella que se utilizará para construir los modelos y consta de n=85 unidades de análisis que corresponden a individuos de ambos sexos que se distribuyen de la siguiente forma (Tabla 14).

Tabla 14. Distribución de los individuos por sexo.

Sexo	n	%
Masculino	52	61.18
Femenino	33	38.82
Total	**85**	**100**

Las variables que se incorporarán como predictoras en los modelos tienen las siguientes características (Tabla 15):

Tabla 15. Estadísticas descriptivas Muestra de Entrenamiento.

	Sexo											
	Masculino						Femenino					
Variable	n	%	Mín - Máx	Media	D. S.	Mediana	n	%	Mín - Máx	Media	D. S.	Mediana
Longitud Máxima Clavícula Derecha	49	62.03	134.4 - 170.1	150.56	7.48	150.6	30	37.97	126.3 - 152.2	137.3	6.94	136.85
Longitud Máxima Clavícula Izquierda	42	62.69	134.2 - 171.5	152.21	7.48	153.3	25	37.31	127.3 - 151.5	138.59	6.51	138.6
Circunferencia Clavícula Derecha	50	60.98	30 - 46	38.24	3.62	37	32	39.02	22 - 39	32.5	3.93	32.5
Circunferencia Clavícula Izquierda	45	60.81	30 - 47	36.76	3.59	37	29	39.19	24 - 37	31.59	2.88	31

Las variables predictoras en los sexos femenino y masculino poseen la siguiente distribución (Figura 20)

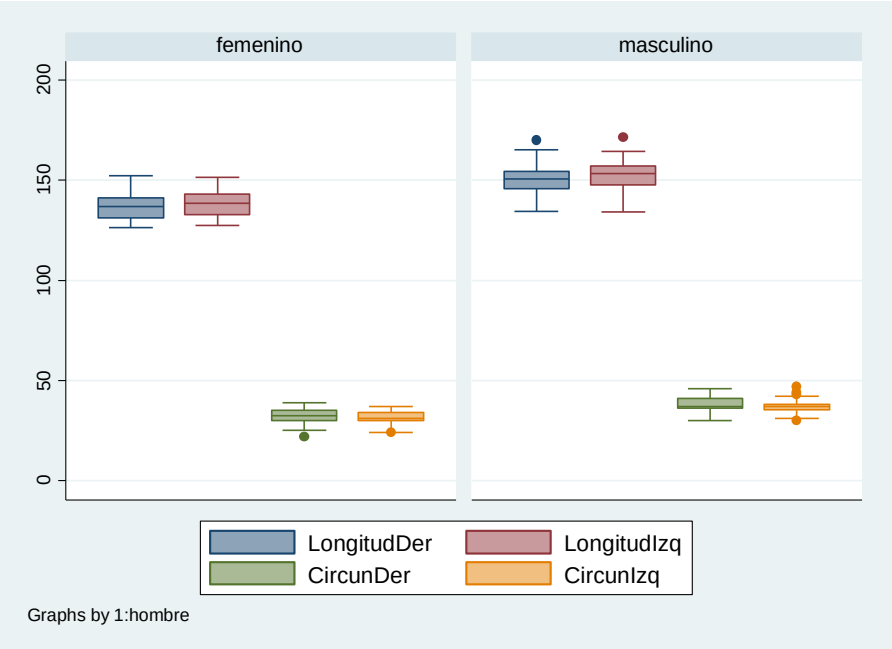


Figura 20. Gráficos de caja y bigotes con la distribución de las variables predictoras en ambos sexos.

Se observa que para ambos sexos existe mayor dispersión en las variables Longitud máxima a diferencia de las variables Circunferencia de eje cuyos datos se hallan más concentrados. En la variable Circunferencia del eje izquierda se identifican varias observaciones atípicas.

6.1.5. Análisis Descriptivo Muestra de Validación.

La muestra de Validación consta de n=37 unidades de análisis que corresponden a individuos masculinos y femeninos cuya distribución se resume en la siguiente tabla (Tabla 16).

Tabla 16. Distribución de la muestra por sexo.

Sexo	n	%
Masculino	24	64.86
Femenino	13	35.14
Total	**37**	**100**

Las variables predictoras muestran las siguientes características (Tabla 17 y Figura 21):

Tabla 17. Estadísticas descriptivas Muestra de Validación.

Variable	Sexo											
	Masculino						Femenino					
	n	%	Min - Máx	Media	D. S.	Mediana	n	%	Mín - Máx	Media	D. S.	Mediana
Longitud Máxima Clavícula Derecha	22	66.67	145.5 - 164.6	152.89	5.36	152.1	11	33.33	123.7 - 167.7	140.5	12.14	141.9
Longitud Máxima Clavícula Izquierda	22	66.67	144.3 - 167.8	155.14	6.69	154.8	11	33.33	127.3 - 160.7	139.37	9.15	138.5
Circunferencia Clavícula Derecha	23	67.65	33 - 49	39.83	4.46	40	11	32.35	27 - 40	33	3.35	33
Circunferencia Clavícula Izquierda	22	64.71	34 - 45	37.82	3.45	37.5	12	35.29	28 - 38	31.83	3.09	31.5

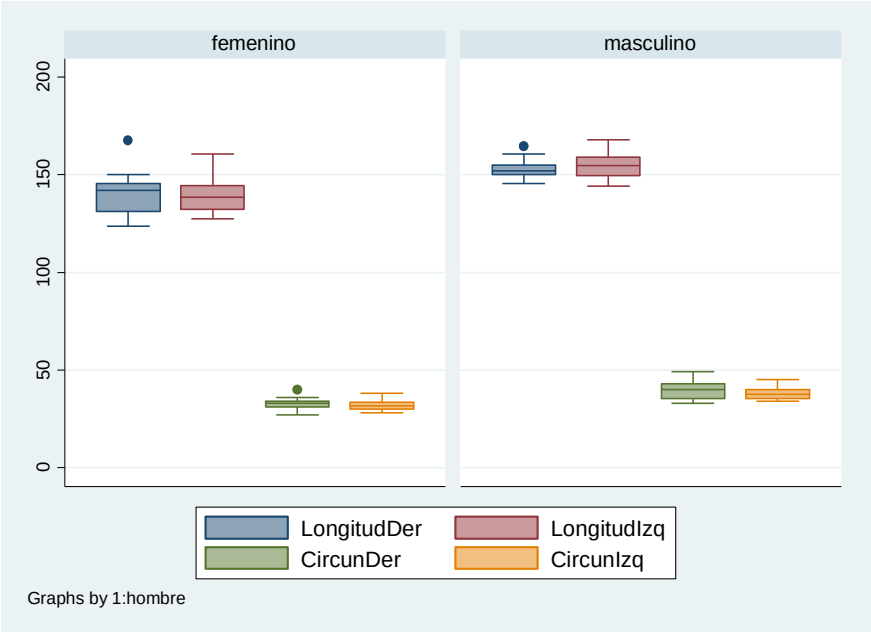


Figura 21. Gráficos de caja y bigotes con la distribución de las variables predictoras por sexo.

6.2. Aplicación de Análisis de Regresión Logística

Se aplicó análisis de Regresión Logística a la muestra de Entrenamiento compuesta por n=85 individuos, usando el programa estadístico Stata 13. Los resultados se exponen a continuación (ver Tabla 18 y 19).

Tabla 18. Resultados del modelo de regresión.

Número de Observaciones	63
Log likelihood	-12.34467
LR chi cuadrado (4)	59.04
Prob > chi cuadrado	0.0000
Pseudo R cuadrado	0.7051

Tabla 19. Coeficientes del modelo de regresión.

Sexo	Coeficientes	Error Estándar	Intervalo Confianza 95 %	
LongitudDer	0.2979106	0.1698652	-0.035019	0.6308402
LongitudIzq	0.0781852	0.1505228	-0.216834	0.3732044
CircunDer	0.4820146	0.288245	-0.0829353	1.046964
CircunIzq	0.2780881	0.2619971	-0.2354167	0.7915929
_cons	-80.29005	23.60573	-126.5564	-34.02368

Se observan los coeficientes de la regresión estimados por el método iterativo de Newton-Raphson. Entre la iteración 3 y 4 la diferencia es mínima, siendo posible indicar que en 4 iteraciones se logró estimar los coeficientes que más verosímilmente pueden haber producido los valores de la variable dependiente (ver Imagen 1 en Anexo).

El modelo utilizó 63 observaciones. Por otra parte, la prueba estadística de significación del modelo basada en Ji- cuadrado indica que la relación entre los coeficientes y la probabilidad de ser un individuo masculino es estadísticamente significativa.

Tabla 20. Resultados del modelo de regresión mostrando odds ratio.

Sexo	Odds Ratio	Error Estándar	Intervalo Confianza 95 %	
LongitudDer	1.347041	0.2288154	0.9655871	1.879189
LongitudIzq	1.081323	0.1627637	0.8050636	1.452381
CircunDer	1.619333	0.4667648	0.9204107	2.84899
CircunIzq	1.320603	0.345994	0.7902415	2.206909
_cons	1.35e-35	3.19e-34	1.09e-55	1.67e-15

En la Tabla 20 se observan los cocientes de razones (odds ratio), sus errores estándar e intervalos de confianza. Estos valores expresan cuanto varía la razón de ocurrencia del suceso (ser individuo masculino) en función del cambio en las variables independientes en una unidad. De acuerdo a esto, si la longitud de la clavícula derecha aumenta en una unidad (1 mm), la razón de ser individuo masculino aumenta 1.35 veces. Se observa que todos los cocientes son mayores que 1 (la razón de ocurrencia aumenta), teniendo, cada variable, un efecto positivo sobre la probabilidad de ocurrencia del suceso.

Luego de calcular los valores predichos por el modelo (con comando “predict p” y cuyos resultados están contenidos en la base de datos), examinamos su capacidad de discriminación con el siguiente gráfico (Figura 22).

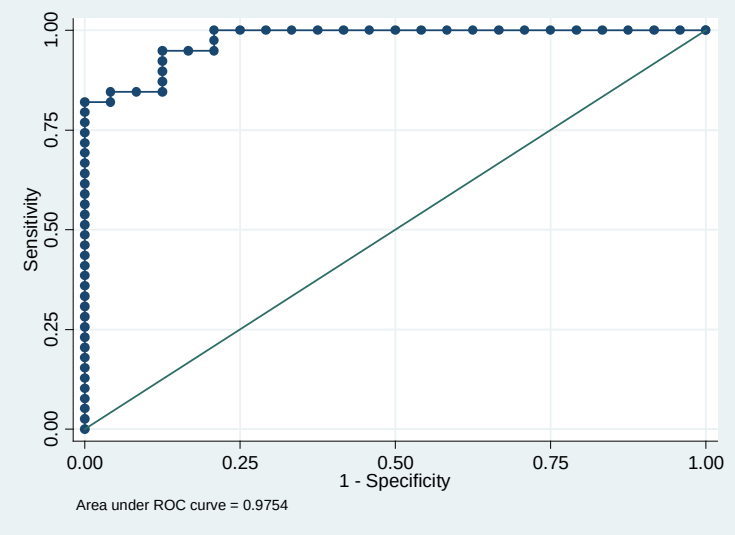


Figura 22. Curva ROC.

Se observa que el área bajo la curva ROC corresponde al valor 0.9754, que se encuentra entre los valores 0.9 y 1 y que de acuerdo a los criterios de Hosmer y Lemeshow [21] se calificaría como discriminación “excelente”.

El punto de corte considerando la sensibilidad y la especificidad se muestra en el siguiente gráfico (Figura 23). Si la discriminación fuese perfecta (sensibilidad 100 % y especificidad 100%), el punto de corte para la discriminación se encontraría en la intersección de ambas curvas y el área bajo la curva azul sería 1, sin embargo en casos reales con buena discriminación, esta área sería menor que 1 pero mayor que 0.5.

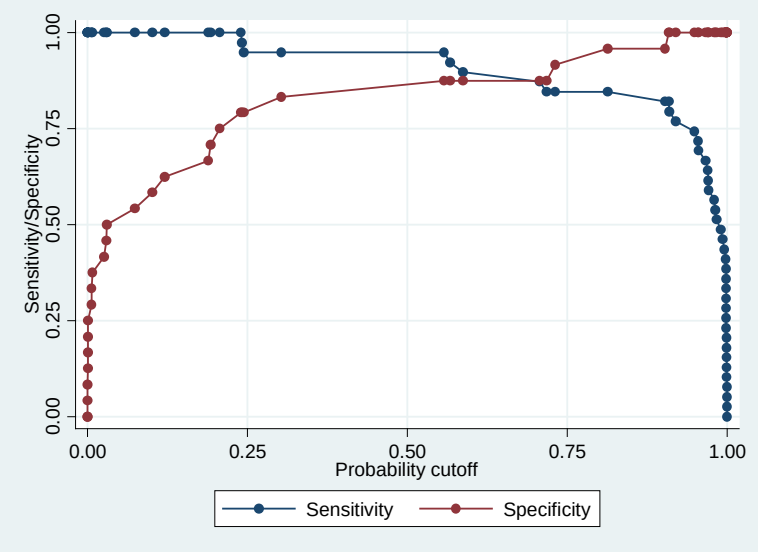


Figura 23. Curvas de sensibilidad y especificidad de la regresión.

Otra forma de observar la discriminación que genera el modelo se muestra en el siguiente gráfico, la línea roja señala el punto de corte entre los individuos masculinos y femeninos (Figura 24). Se puede observar que este punto se aproxima al valor 0.5

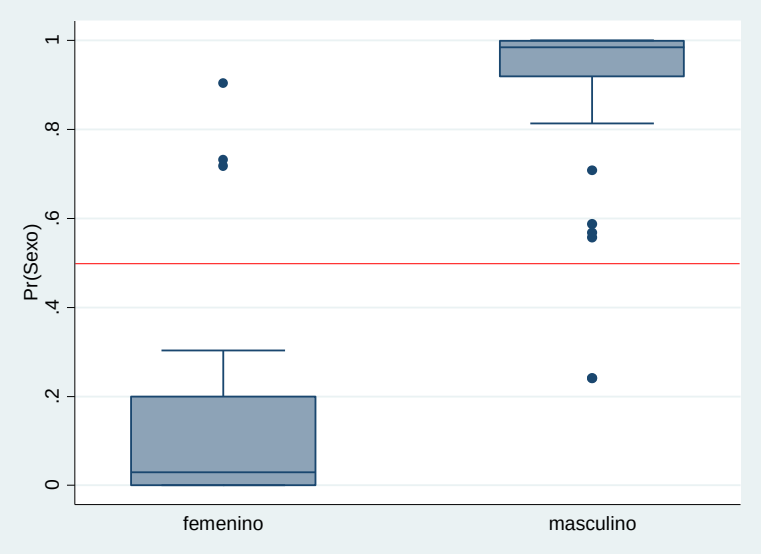


Figura 24. Box plot de los individuos femeninos y masculinos discriminados por la regresión.

En el reporte detallado de la sensibilidad y especificidad del modelo se eligió una probabilidad (punto de corte) de alta sensibilidad y alta especificidad y con un alto porcentaje de casos correctamente clasificados (Tabla 21).

Tabla 21. Puntos de corte según sensibilidad y especificidad del modelo.

Punto de Corte	Sensibilidad	Especificidad	Correctamente Clasificados	LR +	LR -
(>= 0.2444...)	94.87%	79.17%	88.89%	4.5538	0.0648
(>= 0.3033...)	94.87%	83.33%	90.48%	5.6923	0.0615
(>= 0.5574...)	94.87%	87.50%	92.06%	7.5897	0.0586
(>= 0.5672...)	92.31%	87.50%	90.48%	7.3846	0.0879
(>=0.587023...)	89.74%	87.50%	88.89%	7.1795	0.1172

Observaciones	Área ROC	Error estándar	Normal Asintótica Intervalo Confianza 95 %	
63	0.9754	0.0146	0.94679	1.00000

Se escogió el punto de corte 0.5574 y se obtuvo el modelo logístico para la variable sexo con un 92.06 % de casos correctamente clasificados (Tabla 22).

Tabla 22. Modelo logístico para la variable respuesta Sexo.

	Cierto		
Clasificado	**Masculino (D)**	**Femenino (~D)**	**Total**
+	37	3	**40**
-	2	21	**23**
Total	**39**	**24**	**63**

Clasificado + si Pr (D) >= 0.5574 **Realidad D definida como Sexo ≠ 0**		
Sensibilidad	Pr (+|D)	94.87%
Especificidad	Pr (-|~D)	87.50%
Valor Predictivo +	Pr (D|+)	92.50%
Valor Predictivo -	Pr (~D|-)	91.30%
Tasa Falso + para cierto ~D	Pr (+|~D)	12.50%
Tasa Falso - para cierto D	Pr (-|D)	5.13%
Tasa Falso + para clasificado +	Pr (~D|+)	7.50%
Tasa Falso - para clasificado -	Pr (D|-)	8.70%
Correctamente clasificados		92.06%

La regresión logística trabaja la ocurrencia de un evento como el resultado de una combinación lineal de variables. Para la muestra analizada tenemos como resultado del modelo:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * LongitudDer + \beta_2 * LongituIzq + \beta_3 * CircunDer + \beta_4 * CircunIzq$$

Si reemplazamos en la ecuación por los valores correspondientes tenemos:

$$0 = -80.29 + 0.297 * LongitudDer + 0.078 * LongitudIzq + 0.482 * CircunDer + 0.278 * CircunIzq - 0.23$$

Reemplazando las variables predictoras por las dimensiones de clavícula obtenidas de un individuo, un resultado mayor que 0 (positivo) clasificará el sexo del sujeto como masculino, mientras que si el valor obtenido es menor que 0 (negativo), el individuo se clasificará como femenino.

El modelo ajustado fue empleado para predecir la probabilidad de ser clasificado como individuo masculino o femenino utilizando la base de Validación. Para esto se generó la variable Discriminación para definir los casos masculinos y femeninos de acuerdo al modelo logístico generado (ver Tabla 23).

Tabla 23. Sensibilidad y Especificidad del modelo

Clasificación Real	Predicho 0	Predicho 1	Total Casos
0	8 88.88 %	1 11.11 %	9 99.99 %
1	1 5 %	19 95 %	20 100 %

Se clasificaron los individuos en cuatro grupos:

1. Formado por los individuos femeninos que fueron predichos como femeninos.
2. Compuesto por aquellos individuos femeninos que fueron predichos como masculinos.
3. Formado por los individuos masculinos predichos como individuos femeninos.
4. Compuesto por los individuos masculinos predichos como masculinos.

6.3. Aplicación de MARS

Se aplicó análisis MARS a la base de entrenamiento generando dos modelos uno sin interacción y otro con dos interacciones. La variable respuesta corresponde a Sexo codificado con los valores 0 (femenino) y 1 (masculino), mientras que las variables predictoras corresponden a las dimensiones de clavícula.

Los modelos generados son:

Modelo 1: 15 funciones base, sin observaciones entre nodos y sin interacciones (modelo aditivo, ver Tabla 24)

Modelo 2: 15 funciones base, sin observaciones entre nodos y dos interacciones (modelo multiplicativo, ver Tabla 25)

Tabla 24. Descripción del modelo 1

Número Funciones Base	Número Interacciones	Funciones Base	R^2 Ajustado	GCV
4	0	Funciones base que ingresaron al modelo: FB1=(LongitudDer ne . [3]) FB3=max (0, LongitudDer – 132.4) * FB1 FB9=(CircunDer ne .) ; FB11=max (0, CircunDer – 31) * FB9	0.58130	0.11514

[3] La expresión “ne .” se refiere a “no faltante” . “Que es una variable dummy 0/1 que nunca faltará. MARS siempre maneja los valores faltantes usando tales indicadores de valores perdidos”. Dan Steinberg, comunicación personal 24 de Marzo 2017

		FB13=(LongitudIzq ne .) FB15=max (0, LongitudIzq – 136) * FB13 <u>Ecuación del modelo como una combinación lineal de funciones base:</u> Y= 0.601376 + 0.0223885 * FB3 – 0.633923 * FB9 + 0.0497378 * FB11 + 0.00937185 * FB15 <u>Reemplazando en funciones base:</u> Y= 0.601376 + 0.0223885 * max (0, LongitudDer–132.4) * (LongitudDer ne .) – 0.633923 * (CircunDer ne .) + 0.0497378 * max (0, CircunDer – 31) * (CircunDer ne .) + 0.00937185 * max (0, LongitudIzq – 136) * (LongitudIzq ne .)		

Tabla 25: Descripción del Modelo 2.

Número Funciones Base	**Número Interacciones**	**Funciones Base**	**R^2 Ajustado**	**GCV**
3	2	<u>Funciones base que ingresaron al modelo:</u> FB1=(LongitudDer ne .) FB3= max(0, LongitudDer – 132.4) * FB1 FB5=(CircunIzq ne .) FB7=max (0, CircunIzq – 31) * FB5 FB9=(CircunIzq ne .) * FB3 FB11=max (0, CircunIzq – 33) * FB9	0.57168	0.11831

		<u>Ecuación del modelo como una combinación lineal de funciones base:</u> Y= - 0.0261988 + 0.0350964 * FB3 + 0.0893738 * FB7 - 0.00355049 * FB11 <u>Reemplazando en funciones base:</u> Y= - 0.0261988 + 0.0350964 * max (0, LongitudDer – 132.4) * (LongitudDer ne .) + 0.0893738 * max (0, CircunIzq – 31) * (CircunIzq ne .) - 0.00355049 * max (0, CircunIzq – 33) * (CircunIzq ne .) * max (0, LongitudDer – 132.4) * (LongitudDer ne .)		

Se observa que entre los dos modelos, el menor valor para la GCV (validación cruzada generalizada) lo posee el Modelo 1 con 0.11514 contra 0.11831 del Modelo 2. Por otra parte el mayor valor de R^2 ajustado lo presenta el Modelo 1 con un valor de 0.58130 contra 0.57168 del Modelo 2.

Se presenta el orden de importancia de las variables para cada modelo (ver Tablas 26 y 27).

Tabla 26: Importancia de las Variables Modelo 1

Variable	Puntaje
CircunDer_mis[4]	100
CircunDer	99.89
LongitudDer	90.26
LongitudDer_mis	90.26
LongitudIzq_mis	13.96
LongitudIzq	13.96

[4] Las variables que contienen la expresión "_mis" se refieren a aquellas creadas por MARS, después de haber manipulado los datos faltantes.

CircunIzq_mis	0
CircunIzq	0

Tabla 27: Importancia de las Variables en el Modelo 2

Variable	Puntaje
LongitudDer	100
LongitudDer_mis	100
CircunIzq_mis	75.07
CircunIzq	75.07
LongitudIzq	0
LongitudIzq_mis	0
CircunDer_mis	0
CircunDer	0

Se observa que cada modelo jerarquiza de manera distinta las variables. Por una parte para el Modelo 1 la variable Circunferencia del eje de clavícula derecha incluyendo sus datos faltantes ocupa el primer lugar de la jerarquía lo que se opone al Modelo 2, el cual descarta completamente esta variable. De la misma forma, la variable Circunferencia del eje de clavícula derecha sin datos faltantes que en el Modelo 1 ocupa el segundo lugar de la jerarquía, en el Modelo 2 no es considerada. Por otra parte, las variables Longitud máxima de clavícula sin datos faltantes y la misma dimensión con datos faltantes en el Modelo 1 poseen el mismo puntaje (90.26), mientras que en el Modelo 2, encabezan la jerarquía como las variables más relevantes. En cuanto a las variables Longitud máxima de clavícula izquierda con y sin datos faltantes, éstas tienen poca relevancia para el Modelo 1 y nula importancia para el Modelo 2, que no las considera. Finalmente, las variables Circunferencia del eje de clavícula izquierda con y sin valores faltantes están ausentes en el Modelo 1, mientras que para el Modelo 2 ocupan el segundo lugar en la jerarquía.

6.3.1 Indicadores de Bondad de Ajuste.

Para determinar qué modelo es el mejor se examinarán los indicadores de bondad de ajuste de cada uno: valor de GCV y de R^2 ajustado, junto a la sensibilidad y especificidad, además del porcentaje de casos correctamente clasificados. Los dos primeros valores fueron expuestos en las tablas de descripción de los modelos.

Para calcular la sensibilidad y la especificidad de cada modelo generado por MARS, se empleó la base de datos llamada Validación (compuesta por los 37 individuos restantes de la selección aleatoria de individuos inicial) y se procedió de la siguiente manera:

Para el Modelo 1 y utilizando el programa estadístico Stata 14, se generó la variable Y (respuesta o dependiente) que es el resultado de la ecuación del modelo como una combinación lineal de funciones base. Esta variable posee un rango cuyo valor es 1.672301 y que corresponde a la diferencia entre su valor máximo (1.639754) y el valor mínimo (-0.032547). Luego, el rango se multiplicó por la proporción de individuos masculinos reales (64.86%), obteniéndose el punto de corte de la probabilidad de ser evento (ser individuo masculino) dado por el valor 1.0846544.

Así, cuando el valor predicho es menor que 1.0846544, la unidad de observación se clasifica como femenino y cuando el valor predicho es mayor que el punto de corte, el individuo es clasificado como masculino.

Para el Modelo 2, se procedió de la misma forma generándose la variable Y (2) cuyo rango corresponde al valor 2.2458888, diferencia entre los valores 2.21969 (máximo) y -0.0261988 (mínimo). El punto de corte establecido para este modelo fue 1.4566835, por lo que un valor predicho menor a este valor, clasifica como femenino y aquellos mayores al punto de corte, como masculinos.

De lo anterior, se obtuvieron la sensibilidad (hombres predichos por el modelo como masculinos) y la especificidad (mujeres predichas por el modelo como no eventos o femeninos). Estos indicadores se exponen en las siguientes tablas (ver Tablas 28 y 29).

Tabla 28. Clasificación Modelo 1

Clasificación real	Predicho 0	Predicho 1	Total
0	12	1	13
	92.31 %	7.69 %	100 %
1	14	10	24
	58.33 %	41.67 %	100 %

Tabla 29. Clasificación Modelo 2

Clasificación real	Predicho 0	Predicho 1	Total
0	12	1	13
	92.31 %	7.69 %	100 %
1	17	7	24
	70.83 %	29.17 %	100 %

En la Tabla 30 se muestra un resumen de los indicadores de bondad de ajuste de los modelos.

Tabla 30. Resumen de indicadores de bondad de ajuste de ambos modelos

Modelo	R2 ajustado	GCV	Porcentaje clasificación correcta	Sensibilidad (%)	Especificidad (%)
1	0.5813	0.11514	89.19	41.67	92.31
2	0.57168	0.11831	51.35	29.17	92.31

El mejor modelo es aquel que registra el mayor valor de R^2 ajustado, el menor valor de la GCV, los mayores porcentajes de correcta clasificación, sensibilidad y especificidad. De

acuerdo a estos criterios el modelo que mejor ajusta las variables en estudio corresponde al Modelo1.

6.3.2 Análisis Modelo 1

En la siguiente tabla (ver Tabla 31) se muestran indicadores de las 4 funciones base que contiene el modelo ajustado, como la desviación estándar, el costo de omisión que se refiere al costo de pérdida de ajuste del modelo si la función referida es eliminada del modelo, número de parámetros efectivos o grados de libertad de la función base y por último las variables que entran en el modelo.

Tabla 31. Resultados Anova.

Función	Desviación Estándar	Costo de omisión	N° de Parámetros efectivos	Variables
1	0.11697	0.12432	1	CircunDer_mis
2	0.21267	0.14344	1	LongitudDer, LongitudDer_mis
3	0.20294	0.1498	1	CircunDer, CircunDer_mis
4	0.08555	0.11582	1	LongitudIzq, LongitudIzq_mis

Se observa que existe un costo de omisión levemente superior de la función base Circunferencia derecha, que además se apoya con la información contenida en la Tabla 25 sobre la importancia de variables en el Modelo 1 (Tabla 26, pp. 68).

El modelo 1 considera las siguientes funciones base:

- FB1 = (LongitudDer . ne) que corresponde a "Longitud Derecha no faltante"

Esta función base considera el valor cero cuando la variable Longitud derecha es *missing* y un 1 cuando Longitud derecha corresponde a un valor no faltante.

- FB3 = máx. (0, LongitudDer – 132.4) * FB1

La función base 3 considera la variable longitud de clavícula derecha estableciendo como punto de corte para esta medida, 132.4 mm, cuando FB1 es igual a 0, FB3 es 0 y cuando FB1 no es un valor faltante (FB1=1), FB3 = max (0, LongitudDer – 132.4).

- FB9 = (CircunDer ne .) que corresponde a "Circunferencia derecha no faltante"

Esta función base considera el valor cero cuando la variable Circunferencia de clavícula derecha es un valor faltante y un 1 cuando circunferencia de clavícula derecha corresponde a un valor no faltante.

- FB11 = max(0, CircunDer - 31) * FB9

La función base 11 considera la variable circunferencia de clavícula derecha estableciendo como punto de corte 31 mm, cuando FB9 es igual a 0, FB11 es 0 y cuando FB9 no es un valor faltante (FB9=1), FB11 = max (0, CircunDer – 31).

- FB13 = (LongitudIzq ne .)

Esta función base considera el valor cero cuando la longitud de clavícula izquierda es 0 y un 1 cuando longitud de clavícula izquierda corresponde a un valor no faltante.

- FB15 = max (0, LongitudIzq- 136) * FB13

La función base 15 considera la variable longitud de clavícula izquierda estableciendo como punto de corte 136 mm, cuando FB13 es igual a 0, FB15 es 0 y cuando FB13 no es un valor faltante, FB15 = max (0, LongitudIzq- 136).

Ecuación del modelo como una combinación lineal de funciones base:

Sexo predicho = 0.601376 + 0.0223885 * FB3 [(LongitudDer) * (LongitudDer ne .)] – 0.633923 * FB9 (CircunDer ne .) + 0.0497378 * FB11 [(CircunDer) * (CircunDer ne .)] + 0.00937185 * FB15 [(LongitudIzq) * (LongitudIzq ne .)]

Modelo explícito:

P (Sexo masculino = 1) = 0.601376 + 0.0223885 * [máx. (0, LongitudDer – 132.4) * (LongitudDer = 0) o (LongitudDer = 1)] – 0.633923 * (CircunDer = 0) o (CircunDer = 1) +

0.0497378 * [máx. (0, CircunDer – 31) * (CircunDer = 0) o (CircunDer = 1)] + 0.00937185 * [máx. (0, LongitudIzq – 136) * (LongitudIzq = 0) o (LongitudIzq = 1)].

La predicción entregada por el Modelo MARS genera valores próximos a 0 y 1, positivos y negativos. Ya se vio que para obtener la clasificación de los individuos se consideró la proporción de sujetos de sexo masculino presentes en la muestra de Entrenamiento, es decir, 64.86 %, considerando entonces como punto de corte de clasificación de 0.6486 y que correspondería a la prevalencia de individuos masculinos en la muestra de Entrenamiento. Así y para evaluar la pertenencia de la predicción al Sexo masculino, se establece el rango de predicción (al valor predicho máximo se le resta el valor predicho mínimo):

$$\text{Sexo predicho} = \begin{cases} (Masculino) & si\ Valor\ Predicho > Rango\ Predicho * 0.6486 \\ (Femenino) & si\ Valor\ Predicho < Rango\ Predicho * 0.6486 \end{cases}$$

A partir de la modelación MARS y de la obtención de funciones base es posible observar el efecto de las variables predictoras sobre la variable independiente o respuesta por lo que se desprende:

- La Función Base 3: max (0, LongitudDer – 132.4) * FB1 se compone de una variable continua que corresponde al largo máximo de la pieza derecha. El valor 132.4 corresponde al nodo de la función o punto de corte para la variable longitud máxima derecha. Cuando esta variable toma un valor inferior al punto de corte, la función base será igual a 0 (ver Figura 25).

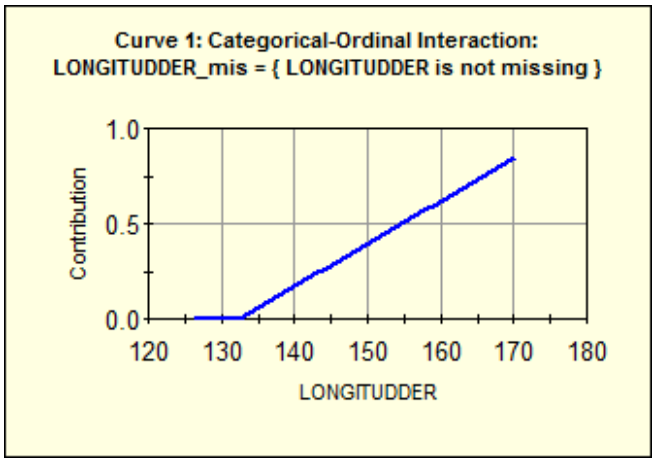


Figura 25. Gráfica de contribución de la variable Longitud máxima de clavícula derecha.

- La Función Base 9: (CircunDer ne .) como ya se mencionó en la descripción del modelo 1, considera el valor cero cuando la variable circunferencia de clavícula derecha es un valor perdido y un 1 cuando circunferencia de clavícula derecha corresponde a un valor no faltante.
- La Función Base 11: max (0, CircunDer - 31) * FB9, considera como punto de corte o nodo de la variable Circunferencia de clavícula derecha el valor 31. Como es posible apreciar en el gráfico (Figura 26) un valor de la variable inferior a 31, resulta un valor 0 para la función base 11.

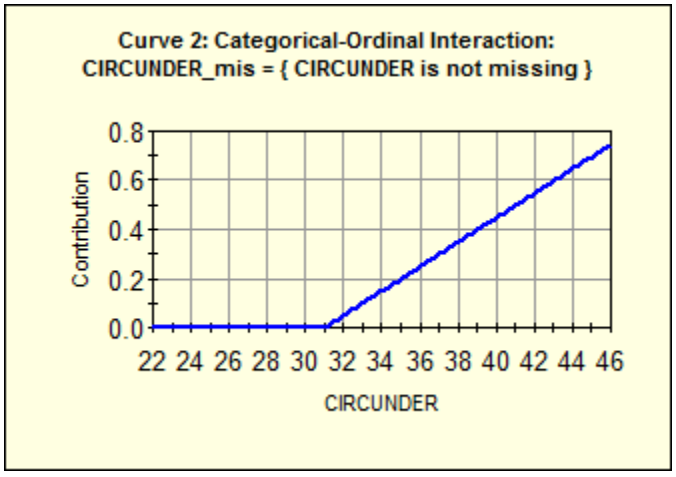


Figura 26. Gráfica de la contribución de la variable Circunferencia derecha de clavícula.

- La Función Base 15: max (0, LongitudIzq - 136) * FB13 contiene el nodo o punto de corte para la variable Longitud máxima de clavícula izquierda que en este caso corresponde al valor 136. Cuando la variable en cuestión toma un valor menor a este punto de corte, la función base es igual a 0 como lo refleja el siguiente gráfico (ver Figura 27).

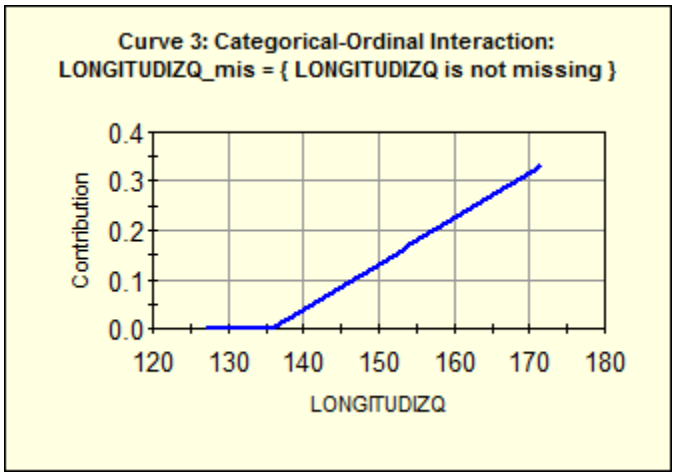


Figura 27. Gráfica de la contribución de la variable Longitud de clavícula izquierda.

6.4. Comparación entre el Modelo de Regresión Logística y Modelo MARS

En la Tabla 31 se exponen los resultados de los análisis de Regresión Logística y MARS modelo 1 (ver Tabla 32).

Tabla 32. Indicadores de Bondad de ajuste de los Modelos de Regresión Logística y MARS.

Método	Sensibilidad	Especificidad	Porcentaje correctas
Regresión Logística	95%	88.88%	93.10%
MARS (Modelo 1)	41.67%	92.31%	59.46%

Los indicadores de bondad de ajuste comparables entre los dos modelos aplicados se concentran principalmente en la correcta predicción del sexo de los individuos de la

muestra denominada Validación. De lo anterior, es posible indicar que el modelo de regresión logística predice de mejor manera los eventos (ser individuo masculino). Adicionalmente, los individuos femeninos son correctamente predichos como no eventos, siendo también un modelo adecuado para este último grupo. Por su parte, el modelo generado por MARS que obtuvo los mayores puntajes en el ítem de indicadores de bondad o Modelo 1, al ser comparado con los resultados de Logit, se presentó como una alternativa insuficiente para predecir los eventos (individuos masculinos), pero obtuvo una alta especificidad o capacidad para predecir los no eventos.

Frente a los hallazgos y recapitulando, el análisis de regresión logística generó un modelo con el 74% de las unidades de observación de la base denominada Entrenamiento. Este modelo estableció, a partir del estadístico de ji-cuadrado para la regresión, que la relación entre sus coeficientes y la probabilidad de ser individuo masculino, es estadísticamente significativa, que la capacidad de discriminación del modelo, dada por el área bajo la curva de ROC (0.9754), es excelente (de acuerdo a la literatura consultada) y que los casos correctamente clasificados corresponden al 92.06 %.

VII. DISCUSIÓN Y CONCLUSIONES

En este estudio se aplicaron dos herramientas de análisis estadístico, Regresión Logística y MARS (Multivariate Adaptive Regression Splines) a una muestra de individuos pertenecientes al universo Patio N° 152 del Cementerio General de Santiago y que correspondieron a todas las unidades de observación posibles para este estudio.

La regresión logística modela la relación entre un conjunto de predictores continuos o categóricos $X = (X_1, \dots, X_p)$ y una variable respuesta dicotómica Y. Produce un modelo para estimar esta respuesta en función de una configuración dada de los predictores X, obteniéndose una probabilidad posterior de ocurrencia del evento y coeficientes asociados a las variables predictoras, los cuales se estiman por máxima verosimilitud. Por otra parte, MARS corresponde a una metodología estadística no paramétrica en la cual la base de datos entrenamiento es particionada en segmentos o ranuras lineales, separadas en regiones con distintas pendientes, siendo estas estructuras conectadas suavemente por piezas polinomiales (funciones base), las cuales a su vez son unidas por nodos que marcan el final de una región y el comienzo de otra. El método tiene la ventaja de efectuar un tratamiento de los datos *missing* generando una nueva variable de tipo binaria.

Se observó que el análisis de regresión logística aplicado a los datos, generó un modelo que ajusta las variables predictoras a la variable respuesta (sexo), de manera satisfactoria en cuanto a sensibilidad, especificidad y porcentaje de individuos correctamente predichos por el modelo. Las cifras para estos indicadores de bondad, son respectivamente: 95 %, 88.88 % y 93.10 %.

En cuanto a la metodología MARS aplicada en este estudio, los indicadores de bondad de ajuste correspondieron a: 41.67 % de sensibilidad, 92.31 % de especificidad y un 59.46 % de los casos correctamente predichos. De esto se desprende que el modelo posee alta especificidad, es decir discrimina de manera satisfactoria a los individuos femeninos, sin embargo, no logra el objetivo para el cual se formuló que es discriminar los eventos, que en este caso corresponde a ser individuo masculino, a partir de las dimensiones de la clavícula seleccionadas.

Frente a los resultados del análisis MARS, es pertinente discutir acerca del tamaño de la muestra empleada. De un n=122, se seleccionó de manera aleatoria una muestra de entrenamiento compuesta por 85 individuos y los restantes constituyeron la muestra de validación con un n=37. Con la muestra de entrenamiento se formuló el modelo que luego se buscó legitimar aplicándolo sobre la segunda muestra obtenida. Es posible que el reducido tamaño de estos grupos no permitiese obtener una buena discriminación de los individuos masculinos.

Los fundamentos que apoyaron la selección de la técnica MARS para ser aplicada en este estudio fueron: por una parte su versatilidad, que permite su aplicación tanto en respuestas de tipo continuas como binarias y porque es pertinente tanto para predictores cuantitativos como cualitativos y por otro lado, por tratarse de una aplicación novedosa en el ámbito de la validación de metodologías para reconstruir el perfil biológico en osamentas humanas.

Por otra parte, el método de Regresión Logística es una alternativa de análisis absolutamente adecuada en presencia de datos *missing*. Permite predecir el resultado de una variable categórica binaria en función de las variables independientes. Constituye una alternativa más robusta que el análisis discriminante clásicamente empleado en el ámbito

de la validación de métodos para estimar el sexo en osamentas. Al respecto, y con relación al análisis discriminante, hay ciertos aspecto que no son considerados en estudios sobre estándares de estimación del sexo, tales como: la relación entre un conjunto de medidas y la probabilidad de ser masculino o femenino es no lineal, la distribución normal multivariante de la variables y la igualdad de las matrices de covarianza entre los dos grupos comparados. La regresión logística, en algunos casos, podría tender a desempeñarse ligeramente mejor en cuanto a sus resultados y a la selección de predictores [27].

El modelo de regresión logística generado y aplicado a la muestra de validación se caracteriza por poseer alta sensibilidad y especificidad, constituyendo una buena herramienta de discriminación del sexo en los individuos analizados.

De acuerdo a la literatura revisada, las variables independientes incorporadas en el modelo, consistentes en dimensiones de clavícula son buenos predictores del sexo en osamentas, sin embargo existe un problema complejo de resolver en cuanto a la representatividad de las muestras que se seleccionan para los fines de validación de este tipo de metodologías. Un ejemplo lo conforma el conjunto de individuos que fue posible de obtener para esta investigación: muestra de osamentas provenientes de un patio de sepulturas temporales del Cementerio General que fueron exhumadas con el fin de reutilizar el espacio de inhumación. El destino final de la mayoría de los individuos extraídos era la cremación, sin embargo, parte de estos restos fueron rescatados para generar una colección de estudio.

El universo al que pertenece la muestra empleada (Patio 152 del Cementerio General) posee rasgos inherentes tales como ausencia de representatividad de todos los grupos etarios y socioeconómicos. Por otra parte, el grado de conservación de las piezas óseas

es bastante variable siendo este un criterio importante en la selección de individuos que conforman una muestra.

Con respecto a uno de los objetivos específicos planteados en este estudio, las variables relevantes en la determinación del sexo por clavícula de acuerdo a la metodología MARS corresponden a las dimensiones de la pieza derecha. Para el modelo Logit, a pesar de que todas las variables tienen un efecto positivo sobre la probabilidad de ocurrencia del suceso o evento, las variables de mayor peso en la estimación (dado por el odds ratio asociado a la característica), también corresponden a aquellas asociadas al lado derecho del esqueleto.

El modelo de regresión logística es muy sensible al problema de la multicolinealidad o alta dependencia existente entre las covariables del modelo. En el punto 6.1.3.3. del capítulo Resultados se efectuó el estudio de la correlación entre variables, el cual expone una fuerte correlación entre las variables Longitudes (derecha e izquierda), por una parte y una alta correlación entre las variables Circunferencias (de ambas clavículas), por otra. Otro problema relevante en el modelo Logit surge de la necesidad de explicar la variable dependiente del modelo con el menor número de regresores posibles.

Considerando los problemas discutidos en los dos últimos párrafos anteriores (importancia de variables en el modelo, multicolinealidad, uso de la menor cantidad de regresores posibles) y rescatando la aplicación forense que se busca dar a los estándares formulados, dados por contextos como: individuos incompletos de los cuales solo se tiene una clavícula disponible para medir, casos de restos mezclados en los cuales no es posible individualizar esqueletos mediante la asociación de piezas; se generaron dos modelos logísticos: uno que considera como variables independientes o predictoras solo

las dimensiones de lado derecho y un segundo modelo ajustado para clavícula izquierda, cuyos predictores son la longitud y la circunferencia de clavícula izquierda (ver ANEXO).

Capítulo VIII: Bibliografía

1. Ubelaker D. *Enterramientos humanos. Excavación, Análisis, Interpretación* (José Luis Prieto, trad.). Donostia: Sociedad de Ciencias Aranzadi; 2007 (Obra originalmente publicada en 1984).

2. Ubelaker D. (Ed.). *Forensic Science: Current Issues, Future Directions*. Chichester, UK: Wiley Blackwell; 2013.

3. Christensen A, Passalacqua N, Bartelink E. *Forensic Anthropology Current Methods and Practice*. Academic Press, Elsevier; 2014.

4. Kamdi A, Gayatri, Sherke A, Krishnaiah, Chaitanya K. *Morphometric Parameters and Sex Determination of Clavicle in Telangana Region*. J. Dent. Med. Sci. 2014; 13(10): 01-05.

5. Albanese J. *A Method for Estimating Sex Using the Clavicle, Humerus, Radius, and Ulna.* J. Forensic Sci. 2013; 58(6): 1413-1419.

6. Jit I, Singh S. *The Sexing of the Adult Clavicles*. Ind. Jour. Med. Res. 1966; 54(6): 551-571.

7. Papaioannou V, Kranioti E, Joveneaux P, Nathena D, Michalodimitrakis M. *Sexual dimorphism of the scapula and the clavicle in contemporary Greek population*: *Applications in forensic identification*. Forensic Sci Int. 2012; 217: 231.e1-231.e7.

8. Králík M, Urbanová P, Wagenknechtová M. *Sex assessment using clavicle measurements: Inter-and intra-population comparisions*. Forensic. Sci. Int. 2014; 234: 181.e1-181.e15.

9. Krenzer U. *Compendio de Métodos antropológico forenses para la reconstrucción del perfil osteo-biológico*. Tomo II. 1ª ed. Guatemala: Centro de Análisis Forense y Ciencias Aplicadas (CAFCA); 2006

10. Sofaer J. The Body as Material Culture. A Theoretical Osteoarchaeology. New York: Cambridge University Press; 2006.

11. Cesani M, Sardi M, Colantonio S, Avena S. Líneas de investigación actuales de la Antropología Biológica Argentina. Revista argentina de antropología biológica. 2014; 16(1): 31-37.

12. Kottak C. Antropología Cultural.14ª ed. (Víctor Campos Olguín, trad.). México: Mc Graw-Hills; 2011.

13. Alcina M., Rissech C, Clavero A, Turbón D. *Sexual dimorphism of the clavicle in a modern Spanish sample.* Eur J Anat*.* 2015; 19(1): 73-83.

14. Iscan M. *Forensic Anthropology of sex and body size*. Forensic Sci. Int. 2005; 147(2): 107-112.

15. Garrido C, Thompson T, Campbell A. *Parámetros métricos para la determinación de sexo en restos esqueletales chilenos modernos.* Chungará, Revista de Antropología Chilena*.* 2014; 46(2): 285-293.

16. Scott J, Freese J. *Regression Models for Categorical Dependent Variables Using Stata*. Texas: Stata Press; 2001.

17. Medina M. Modelos de Elección Discreta. 2003 .Obtenida el 04 de Noviembre de 2015, de http://www.uam.es/personal_pdi/economicas/eva/pdf/logit.pdf/

18. Friedman J. Multivariate Adaptive Regression Splines. *Ann. Statist*. 1991; 19(1): 1-67.

19. Silva C, Pérez P, Trier A. *Statistical modelling and Prediction of Atmospheric pollution by Particulate material: two nonparametric approaches*. Environmetrics. 2001; 12: 147-159.

20. Oyarzún A. *Hacia una Nueva Propuesta en la Selección de las Familias del Programa Chile Solidario: Aplicación de Multivariate Adaptive Regression Splines (MARS) y Análisis basado en Distancias (DB)* [Tesis de magíster]. Santiago, Chile: Escuela de Salud Pública, Universidad de Chile; 2006.

21. Hosmer D, Lemeshow S. *Applied Logistic Regression*. 2nd ed. USA: John Wiley & Sons, Inc; 2000

22. Escobar M, Fernández E, Bernardi F. *Análisis de Datos con Stata.* Cuadernos Metodológicos N° 45. 2ª. ed. Madrid: Centro de Investigaciones Metodológicas; 2012

23. Zhang W, Goh A. *Multivariate adaptive regression splines and neural network models for prediction of pile drivability*. Geoscience Frontiers*.* 2016: 7: 45-52.

24. Oduro S, Metia S, Duc H, Hong G, Ha Q. Multivariate adaptive regression splines models for vehicular emission prediction. *Visualization in Engineering*. 2015; 3(13): 1-12.

25. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* 2nd ed. California: Springer; 2009.

26. Pineda E, de Alvarado E. *Metodología de la Investigación*. 3ª ed. Washington: Organización Panamericana de la Salud; 2008.

27. Santos F, Guyomarc'h P, Bruzek J. *Statistical sex determination from craniometrics: Comparison of linear discriminant analysis, logistic regression, and support vector machines*. Forensic Sci Int. 2014; 245: 204.e1-204.e8.

ANEXO.

Imagen 1. Resultados análisis de regresión con muestra Entrenamiento (resultados con Stata 13).

```
. logit Sexo LongitudDer LongitudIzq CircunDer CircunIzq

Iteration 0:   log likelihood = -41.865292
Iteration 1:   log likelihood = -14.589952
Iteration 2:   log likelihood = -12.698207
Iteration 3:   log likelihood =  -12.34755
Iteration 4:   log likelihood = -12.344671
Iteration 5:   log likelihood =  -12.34467

Logistic regression                               Number of obs   =         63
                                                  LR chi2(4)      =      59.04
                                                  Prob > chi2     =     0.0000
Log likelihood =  -12.34467                       Pseudo R2       =     0.7051

------------------------------------------------------------------------------
        Sexo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 LongitudDer |   .2979106   .1698652     1.75   0.079     -.035019    .6308402
 LongitudIzq |   .0781852   .1505228     0.52   0.603     -.216834    .3732044
   CircunDer |   .4820146    .288245     1.67   0.094    -.0829353    1.046964
   CircunIzq |   .2780881   .2619971     1.06   0.289    -.2354167    .7915929
       _cons |  -80.29005   23.60573    -3.40   0.001    -126.5564   -34.02368
------------------------------------------------------------------------------
```

El modelo anterior corresponde al formulado con dimensiones de ambas clavículas. Sin embargo, con el objeto de aplicar este estándar a contextos forenses en los cuales se recuperan esqueletos muy incompletos, se generaron dos modelos: uno para clavícula derecha (cuyas variables predictoras son Longitud derecha y Circunferencia derecha) y

otro para clavícula izquierda (que considera como predictoras Longitud izquierda y Circunferencia izquierda). De cada modelo se exponen sensibilidad y especificidad.

Imagen 2. Modelo logístico formulado con variables de clavícula derecha.

```
. logit Sexo LongitudDer CircunDer

Iteration 0:   log likelihood = -51.969719
Iteration 1:   log likelihood = -16.233115
Iteration 2:   log likelihood = -14.671895
Iteration 3:   log likelihood = -14.572336
Iteration 4:   log likelihood = -14.572221
Iteration 5:   log likelihood = -14.572221

Logistic regression                             Number of obs   =         78
                                                LR chi2(2)      =      74.79
                                                Prob > chi2     =     0.0000
Log likelihood = -14.572221                     Pseudo R2       =     0.7196

------------------------------------------------------------------------------
        Sexo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 LongitudDer |   .3673148   .1025357     3.58   0.000     .1663486     .568281
   CircunDer |   .7702876   .2468666     3.12   0.002     .2864381    1.254137
       _cons |  -79.76803   21.77508    -3.66   0.000    -122.4464   -37.08966
------------------------------------------------------------------------------
```

En 4 iteraciones se estimaron los coeficientes que más verosímilmente produjeron los resultados de la variable dependiente. Ambas variables son significativas en el modelo. Posteriormente se aplicó el modelo generado (a partir de la base de datos Entrenamiento) a la base Validación para generar los indicadores de sensibilidad y especificidad respectivos (ver Tabla 1).

Tabla 1. Sensibilidad y Especificidad del modelo

Clasificación Real	Predicho 0	Predicho 1	Total Casos
0	10 90.91 %	1 9.09 %	11 100 %
1	1 4.76 %	20 95.24 %	21 100 %

En la Tabla 1 se observa alta sensibilidad, dada por un 95.24 % de los individuos estimados por el modelo como masculinos que efectivamente son masculinos y por otra parte se registra alta especificidad dada por un 90.91 % de los individuos estimados como no eventos (individuos femeninos) cuyo sexo documentado es femenino.

Imagen 3. Modelo logístico para clavícula izquierda.

```
. logit Sexo LongitudIzq CircunIzq

Iteration 0:   log likelihood = -43.788863
Iteration 1:   log likelihood = -17.892415
Iteration 2:   log likelihood = -16.108535
Iteration 3:   log likelihood = -16.045819
Iteration 4:   log likelihood =   -16.0458
Iteration 5:   log likelihood =   -16.0458

Logistic regression                               Number of obs   =         66
                                                  LR chi2(2)      =      55.49
                                                  Prob > chi2     =     0.0000
Log likelihood =   -16.0458                       Pseudo R2       =     0.6336
```

Sexo	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LongitudIzq	.2825558	.0828478	3.41	0.001	.1201771	.4449344
CircunIzq	.5746444	.1937887	2.97	0.003	.1948255	.9544633
_cons	-60.13923	16.12281	-3.73	0.000	-91.73935	-28.53911

El modelo logístico para clavícula izquierda detallado en la Imagen 3 muestra que el modelo es significativo de acuerdo a la prueba estadística de significación basada en el Ji-cuadrado, lo cual en otros términos significa que la relación entre los coeficientes del modelo y la probabilidad de ser individuo masculino es estadísticamente significativa. Adicionalmente, se observa que ambas variables independientes en el modelo son significativas, dado por los valores de la columna P> | z | para las dimensiones de clavícula izquierda.

Con el objetivo de legitimar el modelo generado (con la base Entrenamiento), se aplicó éste a la base Validación y se obtuvieron la sensibilidad y especificidad respectiva (ver Tabla 2).

Tabla 2. Sensibilidad y Especificidad del modelo

Clasificación	Predicho	Predicho	Total

Real	0	1	Casos
0	9 81.82 %	2 18.18 %	11 100 %
1	1 4.76 %	20 95.24 %	21 100 %

Se alcanzó la misma sensibilidad que la registrada en el modelo logístico para clavícula izquierda, dada por un 95.24 % de los individuos predichos por el modelo como masculinos, cuyo sexo documentado es masculino, mientras que la especificidad obtenida fue inferior que en el modelo anterior, con un 81.82 % de los individuos clasificados por el modelo como eventos, cuyo sexo es masculino.