UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Regresión logística y árboles de clasificación

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

M A T E M Á T I C A S

P R E S E N T A:

Ricardo Samaniego de la Fuente

DIRECTORA DE TESIS:
Mat. Margarita Elvira Chávez Cano

Ciudad Universitaria, D.F. 2014

"El hombre es aquello que él hace con lo que hicieron de él."

Jean-Paul Sartre

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Abstract

Facultad de Ciencias
Universidad Nacional Autónoma de México

por Ricardo Samaniego de la Fuente

La regresón logística es un modelo estadístico utilizado frecuentemente cuando se busca ajustar una variable respuesta binaria. El objetivo de esta tésis es hacer una breve introducción a los modelos de regresión logística y a los árboles de clasificación, para después presentar el algoritmo LOTUS, que presenta una forma en la que se pueden combinar ambos. En el primer capítulo se presentan, sin intención de hacer un análisis exhaustivo, algunas de las ideas centrales necesarias para construir modelos de regresión logística simples, así como las motivaciones que llevan al desarrollo de la forma específica del modelo. En el segundo capítulo se presentan modelos con estructura de árboles de clasificación y regresión, que se han desarrollado mediante el algoritmo CART (Breiman, el. al. 1984) como una alternativa a los modelos logísticos. Mediante particiones del espacio muestral se logra clasificar de forma adecuada a los datos. Su estructura gráfica permite una interpretación clara de la relación entre las variables explicativas y la variable respuesta. En el tercer capítulo, se introduce el algoritmo LOTUS, desarrollado por Chan y Loh (Chan y Loh, 2004), en el que se juntan las ideas de los árboles y la regresión logística. El algoritmo LOTUS clasifica y particiona el espacio muestral basándose en el ajuste de modelos logísticos simples. La combinación de ambas técnicas permite evitar algunos de los problemas de éstas, como la dificil interpretación, el problema para determinar cuándo un modelo es adecuado, o el sesgo en las particiones del espacio generado por el algortmo CART. Finalmente, se presentan ejemplos para ilustrar el uso de estas técnicas.

Índice general

Índice de figuras

Índice de cuadros

Para mi familia...

Capítulo 1

Regresión logística

El análisis de regresión es una herramienta estadística utilizada frecuentemente para explicar la relación entre una variable dependiente, y una o más variables explicativas. Puede suceder -lo que no es poco común- que la variable respuesta sea discreta o categórica, en este caso la variable respuesta toma solamente un número finito de posibles valores. En este capítulo nos enfocaremos a revisar el caso en el que toma solo dos posibles valores. Se dice entonces que la variable respuesta es binaria o dicotómica.

Comúnmente los modelos de regresión que más se usan son los modelos lineales. En el caso de variables respuesta dicotómicas se usa la regresión logística. Al modelar un tipo distinto de variable, el ajuste del modelo debe ser también distinto, así como las suposiciones previas. Hay que destacar que, a pesar de las diferencias, las metas de la regresión logistica y lineal son las mismas: construir el modelo que de mejor y más sencilla forma explique el problema a modelar.

Ejemplo
Para ilustrar el uso de la Regresión logística, se comienza por ilustrar el problema con un ejemplo. Se usa el conjunto de datos del Condado de Evans (Evans County Dataset), en el que se presentan los datos de un estudio en el que se siguió a 609 pacientes masculinos durante 7 años, siendo la presencia de síndrome coronario (CHD) la variable de interés. A continuación se presenta una descripción de las variables:

- ID: identificador del individuo
- CHD: presencia (1) o ausencia (0) de síndrome coronario.
- CAT: nivel alto o normal de catecotamina (codificado como 1 o 0)
- AGE: edad del individuo

- CHL: nivel de colesterol
- SMK: variable indicadora de si el paciente fumó alguna vez o no (1 o 0)
- ECG: presencia o ausencia de anormalidades en el electrocardiograma (1 o 0)
- DPB: presión sanguinea diastólica
- SBP: presión sanguinea sistólica
- HPT: indicadora de presión alta
- CH: producto de CAT x HPT
- CC: producto de CAT x CHL

Se presenta a continuación una tabla con las primeras observaciones del conjunto de datos.

Cuadro 1.1: Resumen de variables (Evans County Dataset)

	ID	CHD	CAT	AGE	CHL	SMK	ECG	DPB	SPB	HPT	CH	CC
1	21	0	0	56	270	0	0	80	138	0	0	0
2	31	0	0	43	159	1	0	74	128	0	0	0
3	51	1	1	56	201	1	1	112	164	1	1	201
4	71	0	1	64	179	1	0	100	200	1	1	179
5	74	0	0	49	243	1	0	82	145	0	0	0
6	91	0	0	46	252	1	0	88	142	0	0	0

Cuadro 1.2: Análisis descriptivo de variables (Evans County Dataset)

	AGE	CHL	DPB	SPB
min.	40	94	60	92
1er cuartil	46	184	80	125
media	52	209	90	140
mediana	54	212	91	145
3er cuartil	60	234	100	160
max.	76	357	170	300

Cuadro 1.3: Análisis descriptivo de variables (Evans County Dataset)

	CHD	CAT	SMK	ECG	HPT
1	538	487	222	443	354
0	71	122	387	166	255

El análisis de regresión es una herramienta estadística para investigar y modelar la relación entre variables. En muchas áreas de investigación, el investigador se puede preguntar cuál es la relación entre una variable respuesta categórica (por ejemplo la presencia de

una enfermedad) y una o varias variables explicativas o variables de estudio. Por ejemplo, un médico puede querer explicar la influencia de la edad de los pacientes (AGE) y su estatus de fumadores (SMK) sobre la presencia de síndrome coronario en el corazón (CHD). En este caso, la variable dicotómica o respuesta es CHD, que se representa con un 1 la presencia de la enfermadad, y con 0 su ausencia; las variables independientes o explicativas son AGE y SMK. En este caso, la variable AGE es continua, mientras que la variable SMK es también dicotómica. Se denota a las k variables independientes como X_i, con $i = 1, ..., k$, y a la variable dependiente como Y. Usualmente, se busca algún modelo matemático que ayude a explicar e interpretar la relación entre estas variables.

1.1. Modelo lineal y modelo logístico

En cualquier modelo de regresión el valor de interés es el valor esperado de la variable respuesta Y dados los valores de las variables explicativas X, esto es $E(Y = y|x)$, también llamada la media condicional. La primera diferencia entre el modelo logístico y el lineal es la naturaleza de dicha relación. Supóngase que se tiene un conjunto de n observaciones independientes entre sí Y_i, donde $i = 1, ..., n$. Si el objetivo es predecir el valor de una nueva observación, pero no se cuenta con ninguna variable explicativa, la mejor aproximación sería usar $\bar{Y}$, que es la media de la variable Y, $E(Y)$. Por otro lado, supongase ahora que se cree que con las observaciones de una variable independiente X se puede obtener una mejor aproximación. Entonces lo que se está buscando es la media o esperanza condicionada al valor de X, $E(Y|X = x)$. En regresión lineal, se supone que la media condicional se puede expresar como una transformación lineal de la variable X, $E(Y|X = x) = \beta_0 + \beta_1 x$, lo que implica que el valor de $E(Y|X = x)$ puede encontrarse entre $-\infty$ e ∞. Para variables dicotómicas, Y solo puede tomar los valores 0 y 1, por lo que la media condicional de Y debe estár necesariamente entre 0 y 1. Es por esto que el modelo lineal no servirá para modelar a Y como variable respuesta.

1.1.1. Por qué el modelo logístico?

La función logística, en la que se basa el modelo, tiene las siguientes formas:

$$f(x) = \frac{1}{1 + e^{-z}} \tag{1.1}$$

$$g(x) = \frac{e^{-x}}{1 + e^{-x}} \tag{1.2}$$

La elección de dicha función para modelar el problema es adecuada para nuestro problema, pues cumple con varias propiedades convenientes. El rango de la función va de 0 a 1: cuando los valores de z tienden a $-\infty$, f(x) se va acercando a 0, mientras que cuando crecen a ∞, f(x) se acerca a 1. Nótese que g(x) tiene básicamente la misma forma que f(x), pero invertida. El modelo logístico sirve para explicar una probabilidad o riesgo (en el ejemplo anterior la probabilidad de tener CHD), por lo que el valor se debe encontrar siempre entre cero y uno. Es por esto que una de las razones principales para utilizar el modelo logístico es su rango. Otra propiedad que tiene el modelo y por la que también es adecuado utilizarlo es su forma. En la figura 1.1 presentada a continuación, se puede observar como para valores muy pequeños de z, la función se encuentra muy cerca del cero. Al ir aumentando el valor de z, f(z) va creciendo, hasta que alcanza cierto valor a partir del cual la función se aproxima al uno. El resultado es que la función toma la forma de una S. Si se considera a la variable z como la aportación de las variables independientes para explicar la presencia de la enfermedad, entonces se puede interpretar la forma de S como sigue: para valores muy pequeños de z el riesgo o la probabilidad de tener la enfermedad es muy baja. Al ir creciendo, la probabilidad también lo hará, hasta alcanzar cierto valor a partir del cual la probabilidad de contraer la enfermedad será muy grande, casi uno.

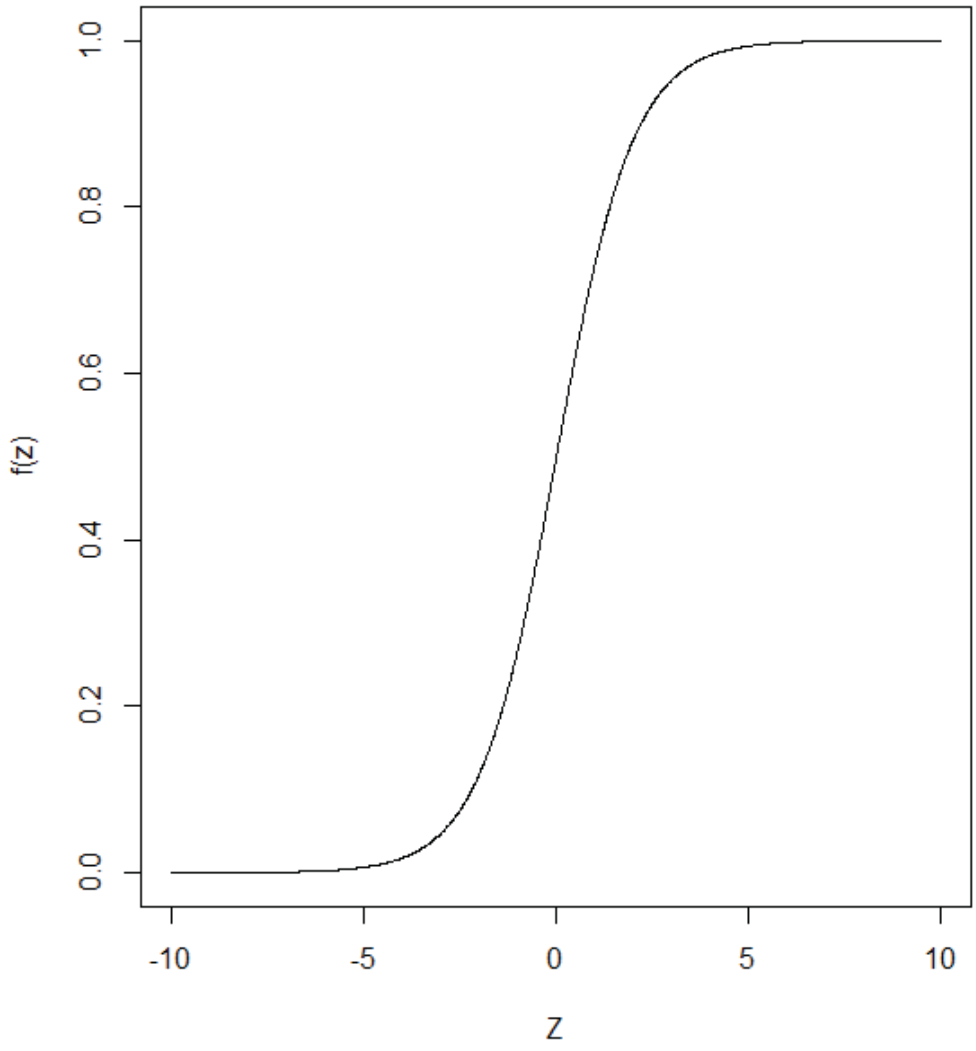


FIGURA 1.1: Función logística

1.1.2. Modelo logístico

Fue mencionado anteriormente que dentro de la función logística, se puede interpretar a z como una combinación de la aportación de las variables independientes X_i's. Se escribe entonces a z como

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

donde las β_i's son parámetros constantes a estimar y las X_i's son las variables explicativas. En esencia, z representa un índice con las aportaciones de las X. La función logística con z vista de esta forma, queda de la siguiente forma:

$$f(z) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k)}}$$

o bien

$$f(z) = \frac{e^{-(\beta_0+\beta_1 X_1+...+\beta_k X_k)}}{1 + e^{-(\beta_0+\beta_1 X_1+...+\beta_k X_k)}}$$

A partir de este momento se denotará a $E(Y|X = x)$ como $\pi(x)$. La forma que se usa del modelo de regresión logística, para el caso de una sola variable independiente, es la siguiente:

$$\pi(x) = \frac{e^{-(\beta_0+\beta_1 X)}}{1 + e^{-(\beta_0+\beta_1 X)}} \tag{1.3}$$

La función *logistic*,

$$P(Y) = \frac{\exp^Y}{1 + \exp^Y} \tag{1.4}$$

tiene varias interpretaciones, pero la concerniente en este caso es aquella que se le da en el uso para la regresión. En este caso P será la probabilidad de un resultado binario, donde se modela a Y como $Y = \alpha + \beta X$, con X una variable independiente y posiblemente continua. En este caso α es el intercepto (el valor de Y cuando X vale cero), mientras que β representa la pendiente, que nos da el cambio en Y por un aumento en X. Originalmente, la función *logistic* era utilizada para describir el cambio de una proporción P con el tiempo t, entonces $Y = \alpha + \beta t$; dado que P(t) crece monotónamente con el tiempo, se puede ver como una curva de crecimiento. El logit de un número P entre cero y uno, se define entonces como

$$logit(P) = log(\frac{P}{1-P}) = log(P) - log(1 - P) \tag{1.5}$$

La función o transformación logistic es la inversa del logit de P. Para interpretar dichas funciones, se toma a P como una probabilidad. Entonces, el número $P/1 - P$ representa los momios, la razón entre la probabilidad de éxito y la probabilidad de fracaso. Parte de la importancia de utilizar la transformación logit yace en que P es lineal cuando se modela como $\alpha + \beta X$, y con rango $(-\infty, \infty)$.

La transformación logit, expresada en términos de $\pi(x)$ se ve de la siguiente manera:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X$$

Es importante darse cuenta que la probabilidad, los momios y la transformación logit son tres formas distintas de expresar lo mismo, ya que a través de transformaciones y sus inversas, se llega de una a cualquiera de las otras.

1.1.3. Distribución condicional de la variable respuesta

Una pregunta importante es cuál es la distribución condicional de la variable dependiente. En la respuesta a esta pregunta reside la segunda diferencia importante con el modelo de regresión lineal, en el cual se supone que una observación de la variable respuesta puede expresarse como $y = E(Y|x) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$. El valor ε es llamado el error, y se puede interpretar como la desviación de la media condicional. Se acostumbra suponer que el error sigue una distribución Normal con media 0 y varianza constante σ^2. De aquí se sigue que el valor de y dado x sigue una distribución Normal con media $E(Y|x)$ y varianza constante. Ahora, cuando la variable es binaria y se expresa el valor de una observación como $y = \pi(x) + e$, el valor e tiene dos posibilidades. Si Y=1, entonces $\varepsilon = 1 - \pi(x)$ con probabilidad $\pi(x)$, mientras que cuando Y=0, $\varepsilon = -\pi(x)$ con probabilidad $1 - \pi(x)$. En este caso, la distribución condicional de la variable respuesta Y es Bernoulli con probabilidad dada por el valor de $\pi(x)$.

Para ilustrar como funciona el modelo logístico, se continúa con el ejemplo del *Evans County Dataset*. Supongase que se tiene para un grupo de n individuos (en este caso n pacientes) las observaciones correspondientes de su edad (AGE) y si fuman o no (SMK). También se tiene la información sobre la presencia o ausencia de la enfermedad (CHD). Se quiere entonces usar la información para modelar la probabilidad de tener la enfermedad ($P(Y = 1)$) dados los valores de AGE y SMK, las variables explicativas. El modelo logístico entonces es aquel en el que modelamos dicha probabilidad utilizando la función anterior.

$$\pi(x) = P(Y = 1|X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}}$$

Un ejemplo del uso que se puede dar al modelo es el siguiente: si se conocieran los parámetros β_i's para un paciente nuevo, se podría calcular la probabilidad de que este tenga la enfermedad a través de la información de su edad y de conocer si fuma. Más adelante se ve como estimar los valores de los parámetros desconocidos usando la información observada. Por ahora se supone que se quiere estudiar la presencia de la

enfermedad CHD, codificada como 1 si el paciente está enfermo y 0 si no lo está. Las variables independientes que serán utilizadas para modelar el comportamiento de CDH son las siguientes: $X_1 = CAT$, el nivel de catecolamina, categorizado como 1 si es alto y 0 si es bajo, $X_2 = AGE$, la edad del paciente, y $X_3 = ECG$, que corresponde al estátus del electrocardiograma (codificado como 1 si es anormal y 0 si es normal) . En este caso se tienen dos variables categóricas y una continua. Recuerdese que los datos consisten de observaciones para 609 pacientes, para los que el objetivo es modelar logísticamente la presencia de la enfermedad. El objetivo entonces es ajustar un modelo: usar los datos observados de los 609 pacientes para estimar los parámetros β_0, β_1 y β_2, en el siguiente modelo:

$$\pi(x) = (Y = 1|X_i's) = \frac{e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3)}}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3)}}$$

1.2. Ajuste del modelo y estimación de los parámetros

Para el caso de un modelo en el que contamos con solo una variable independiente, se tienen n observaciones de la forma (x_i, y_i), con $i = 1, ..., n$. El ajuste del modelo requiere estimar entonces los valores de β_0 y β_1. En regresión lineal se acostumbra hacer el ajuste usando el método de mínimos cuadrados. El método general que se aplica en el caso de la regresión logística es el de máxima verosimilitud. En términos muy generales, la idea detrás del método es obtener estimadores de los parámetros desconocidos que maximicen la probabilidad de obtener los valores observados.

1.2.1. Método de máxima verosimilitud y bondad de ajuste

A continuación se verá como construir la función de verosimilitud. En términos generales, dicha función (de uno o más parámetros desconocidos, en este caso β_0 y β_1) da la probabilidad de obtener un conjunto de observaciones dados los valores de los parámetros desconocidos. Al maximizar la función, se obtienen los estimadores buscados, llamados estimadores máximo verosímiles. Se revisarán algunas de sus propiedades, como su distribución para muestras grandes. Por último se verá como se construyen los intervalos de confianza para los estimadores máximo verosímiles, que serán de utilidad cuando se hable sobre bondad de ajuste. La función de verosimilitud para un parámetro β dadas n observaciones $x_1, ..., x_n$, se define como

$$L(\beta; X) = \prod_{i=1}^{n} f(x_i, \beta) \tag{1.6}$$

El estimador máximo verosímil $\hat{\beta}$ es el valor de β que maximiza a $L(\beta; X)$.

1.2.2. Distribución para muestras grandes

Es razonable suponer que un buen método de estimación proporciona estimadores que mejoran conforme el tamaño de muestra crece. Matemáticamente dicha condición es equivalente a pedir que los estimadores converjan en probabilidad al parámetro real que están estimando conforme n, el tamaño de muestra, tiende a infinito. En estadística a dicha propiedad de los estimadores se le llama consistencia. Por ejemplo, si se supone que se está muestreando sobre una distribución arbitraria con varianza finita σ^2, la media $\bar{X}$ es un estimador consistente de la media μ. En este caso, se puede asegurar por el Teorema del Límite Central, que

$$\frac{(\bar{X} - \mu)\sqrt{n}}{\sigma} \rightarrow N(0, 1)$$

.

Se dice que $\bar{X}$ tiene distribución asintótica $N(\mu,\sigma^2)$.

1.2.3. Teoría para muestras grandes: estimadores máximo verosímiles

Se muestra a continuación que los estimadores máximo verosímiles tienen una distribución asintótica Normal. Se define a la función:

$$I(\theta) = -\int \left(\frac{\partial^2}{\partial\theta^2} log f(x,\theta)\right) f(x,\theta)\,\mathrm{d}x = -E\left(\frac{\partial^2}{\partial\theta^2} log f(X_1, \theta)\right)$$

como la información de Fisher observada. Para una muestra de tamaño n, análogamente:

$$I_n(\theta) = -\int \left(\frac{\partial^2}{\partial\theta^2} log f(X,\theta)\right) f(X,\theta)\,\mathrm{d}X = -E\left(\frac{\partial^2}{\partial\theta^2} log f(X,\theta)\right)$$

Bajo condiciones de regularidad (véase por ejemplo Casella, cap.10), el estimador máximo verosímil $\hat{\theta}$ de θ sigue una distribución asintótica Normal, tal que

$$\hat{\theta} \approx N(\theta, \frac{1}{I_n(\theta)}) \text{cuando n } \to \infty$$

1.2.4. Consistencia de Estimadores máximo verosímiles

Sea $X_1, ...X_n$, una muestra aleatoria de $f(x|\theta)$, y sea $\hat{\theta}$ el estimador máximo verosímil de θ. Para una función continua $\rho(\theta)$ y bajo condiciones de regularidad, para todo $\epsilon > 0$ y para todo θ en Θ se cumple que

$$\lim_{n \to +\infty} P_\theta(|\rho(\hat{\theta}) - \rho(\theta)| \geq \epsilon) = 0$$

es decir, $\rho(\hat{\theta})$ es un estimador consistente de $\rho(\theta)$. el teorema dice que cuando el tamaño de muestra crece, el estimador estará tan cerca como se quiera del valor a estimar con probabilidad tan grande como sea necesario.

1.2.5. Intervalos de confianza basados en los estimadores máximo verosímiles

Para $\hat{\theta}$ un estimador máximo verosímil de θ, se sabe que se aproxima asintóticamente la varianza de $\hat{\theta}$ con la información de Fisher, $I(\theta)$. Para un valor fijo pero arbitrario de θ,

$$\frac{\hat{\theta} - \theta}{\sqrt{\widehat{var(\theta)}}} \to N(0, 1)$$

Por lo que un intervalo de $(1 - \alpha) \times 100\,\%$ de confianza será:

$$\hat{\theta} - z_{1-\alpha/2}\sqrt{\hat{var}(\hat{\theta})}, \hat{\theta} + z_{\alpha/2}\sqrt{\hat{var}(\hat{\theta})}$$

donde $\widehat{var(\theta)} = 1/I(\hat{\theta})$.

1.2.6. Distribución asintótica del cociente de verosimilitudes

Para probar $H_0 : \theta = \theta_0$ contra $H_1 : \theta \neq \theta_0$, y donde $X_1, ..., X_n$ son variables aleatorias i.i.d. de $f(x|\theta)$ y $\hat{\theta}$ es el estimador máximo verosímil de θ, y $f(x|\theta)$ satisface condiciones de regularidad (referirse a Casella, 10.6.2, 2002), entonces bajo H_0 y cuando $n \to \infty$,

$$-2log\lambda(X) \rightarrow \chi^2{}_{(1)} \text{en distribución,}$$

donde $\chi^2{}_{(1)}$ es una variable aleatoria jí-cuadrada con un grado de libertad.

El teorema anterior se extiende para los casos en los que la hipótesis nula $H_0 : \theta \in \Theta_0$ se refiere a un vector de parámetros $\theta \in \mathbb{R}^p$. En este caso, la distribución de $-2log\lambda(X)$ se distibuye jí-cuadrada con grados de libertad equivalentes a la diferencia entre el número de parámetros libres dados por H_0 y el número dado por la hipótesis alternativa (ver Casella, 2002, cap.10).

1.3. Ajuste por el método de máxima verosimilitud

Regresando al problema del ajuste de la regresión, la pregunta que se plantea es la de cómo estimar los parámetros β_0 y β_1. Usualmente en regresión lineal el método utilizado es el de mínimos cuadrados. Si se supone que se tiene un conjunto de n observaciónes (x_i, y_i), entonces

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{ para toda } i = 1, ..., n$$

La suma de los cuadrados de las desviaciones verticales de la linea de regresión es la suma de los cuadrados del error e

$$S = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

.

Los valores β_0 y β_1 se eligen de tal forma que esta suma de cuadrados se minimice. Los estimadores de β_0 y β_1 están dados por la solución al siguiente sistema:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

de donde se obtienen las ecuaciones normales,

$$\sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i = 0$$

$$\sum X_i Y_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0$$

La solución para la pendiente $\hat{\beta}_1$ en las ecuaciones anteriores resulta ser

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

y para $\hat{\beta}_0$,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Cuando la variable respuesta a modelar es dicotómica o binaria, debido a las consideraciones anteriores, cambia la forma de modelar su relación con la o las variables independientes. El modelo que se usa es el de la distribución logística. Se observó ya que en la regresión lineal el error se supone normal $(0, \sigma^2)$. Este no puede ser el caso cuando la variable respuesta es dicotómica. En este caso, se puede expresar el valor de la variable respuesta como $y = \pi(x) + e$. En este caso, el error e puede tomar dos valores: si y=1, $e = 1 - \pi(x)$ con probabilidad $\pi(x)$, mientras que si y=0, $e = -\pi(x)$ con probabilidad $1 - \pi(x)$. Entonces, e sigue una distribución con media cero y varianza $\pi(x)(1 - \pi(x))$. Es decir, el error sigue una distribución binomial con media dada por la esperanza condicional $\pi(x)$. Debido a la forma lineal que toma la transformación logit, y ya que toma posiblemente cualquier valor en los reales, no se puede estimar a los parámetros con el método tradicional de mínimos cuadrados. En regresión lineal, se eligen como estimadores de los parámetros a y b aquellos valores que minimícen la suma de las desviaciónes al cuadrado de los valores observados de Y contra los valores estimados con el modelo. Esto se puede ver como

$$\sum (Y_i - \hat{Y}_i)^2$$

$\pi(x)$ representa la probabilidad condicional de que Y=1 dado el valor de x. Así, $1 - \pi(x)$ arroja la probabilidad condicional de que Y=0 dado x. Supóngase entonces que se tienen n parejas de observaciónes (y_i, x_i). Para aquellas parejas en las que $y_i = 1$, su contribución a la función de verosimilitud será $\pi(x_i)$, mientras que para aquellas en las que $y_i = 0$, ésta será $1 - (\pi(x_i))$. Se puede entonces expresar para toda y_i la contribución a la función de verosimilitud como

$$\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

Suponiendo que las observaciones son independientes, la función de verosimilitud toma la siguiente forma

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

El principio del método de máxima verosimilitud es encontrar los valores de β que maximicen a la función anterior. Lo que se hace equivale a encontrar los valores de los parámetros tales que la probabilidad de haber obtenido dichas observaciones (y_i, x_i) sea máxima. Ahora, si aplicamos logaritmo, se obtiene

$$l(\beta) = ln(L(\beta)) = \sum_{i=1}^{n} (y_i ln(\pi(x_i)) + (1-y_i) ln(1-\pi(x_i)))$$

Para encontrar el valor de β que maximice a $l(\beta)$ basta diferenciar respecto a β e igualar las ecuaciones a cero. Las ecuaciones conocidas como las ecuaciones de verosimilitud toman la siguiente forma:

$$\sum_{i=1}^{n} (y_i - \pi(x_i)) = 0 \tag{1.7}$$

$$\sum_{i=1}^{n} (x_i(y_i - \pi(x_i)) = 0 \tag{1.8}$$

Estas ecuaciones no son lineales en β, por lo que su solución requiere de utilizar métodos numéricos, que consisten en empezar con una solución inicial, e irla corrigiendo iterativamente hasta que el cambio en la solución entre una iteración y la siguiente sea insignificante. Se considera entonces que la solución converge a $\hat{\beta}$ y se toma a dicho $\hat{\beta}$ como el estimador máximo verosímil. De momento no se presenta de forma específica cuáles son los métodos iterativos que se suelen utilizar para estimar los parámetros, pero basta mencionar que obteniéndolos se puede obtener el estimador de $\pi(x_i)$, denotado $\widehat{\pi(x_i)}$, que resulta de evaluar la función $\pi(x_i)$ en $\hat{\beta}$. Este valor nos da un estimado de la probabilidad de que Y=1 dado el valor x_i de x.

Para ilustrar esto se retoma el ejemplo que se ha utilizado. El modelo a ajustar es el siguiente:

$$\pi(x) = P(Y = 1|X_i's) = \frac{1}{1 + e^{-(\beta_0+\beta_1 CAT+\beta_2 AGE+\beta_3 ECG)}}$$

Los valores de los estimadores de β_0, β_1,β_2 y β_3, denotados $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ y $\hat{\beta}_3$ que resultan del ajuste son los siguientes:

- $\hat{\beta}_0 = 3.911$
- $\hat{\beta}_1 = 0.652$
- $\hat{\beta}_2 = 0.029$
- $\hat{\beta}_3 = 0.342$

Por lo tanto, el modelo logístico es el siguiente:

$$\widehat{\pi(x)} = P(\widehat{Y = 1|X_i's}) = \frac{1}{1 + e^{-(3.911+0.652X_1+0.029X_2+0.342X_3)}}$$

Si lo que se quiere es obtener la probabilidad de que un individuo contraiga la enfermedad coronaria, se deben especificar los valores de las variables independientes. Supóngase por ejemplo que un paciente presenta las siguientes observaciones: AGE=40, CAT=1 y ECG=0. Se tiene entonces que

$$\widehat{\pi(1, 40, 0)} = \frac{1}{1 + e^{-(-3.91+0.652+0.029(40))}} = 0.1090$$

La probabilidad o el riesgo del paciente de contraer síndrome coronario es de 0.109.

1.4. Pruebas de significancia

Una vez ajustado el modelo, el siguiente paso es verificar si la relación entre la o las variables independientes y la variable dependiente es significativa. Una forma de aproximarse al problema es comparando entre el modelo que incluye a las variables independientes X y un modelo que no las incluya. Se comparan entonces los valores observados de Y con las predicciones obtenidas mediante los dos modelos anteriores. Cabe destacar que no se está verificando que tan preciso es el ajuste comparandolo con las observaciones, sino comparando entre el modelo que contiene a las variables independientes contra uno en el que no se incluyen.

En regresión lineal la prueba de significancia se hace mediante la elaboración de una

tabla de análisis de varianza (ANOVA), en la que se particiona la suma total de las desviaciones al cuadrado de las observaciones alrededor de su media en dos partes. La primera es la suma de las desviaciones al cuadrado alrededor de la recta de regresión. La segunda es la suma de cuadrados de las desviaciones de las predicciones basadas en el modelo de regresión alrededor de la media de la variable dependiente. La idea es que cuando se quiere minimizar los errores en la predicción, si solo se conoce la variable dependiente, la mejor forma de aproximar las observaciones es mediante la media $\bar{Y}$. En este caso, la suma a minimizar es

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Donde SST es la suma de cuadrados total *(sum of squares total)*. Si se conocen los valores de la variable independiente X, y se supone que con ellos se puede ajustar un modelo para predecir el valor de Y, entonces obtenemos,

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

.

En este caso SSE es la suma de cuadrados de los residuales *(residual sum of squares)*.

En regresión lineal se busca minimizar SSE mediante el método de mínimos cuadrados. El modelo que no contiene a la variable independiente tiene como único parámetro a β_0, y $\hat{\beta}_0 = \bar{Y}$. En este caso $\hat{y}_i = \bar{Y}$ y SSE equivale a la variabilidad total. Al incluir a la variable independiente, cualquier decrecimiento en SSE se deberá a que la pendiente (es decir el coeficiente de la variable independiente) es diferente a cero. Éste cambio en el valor de SSE se debe a la variabilidad debida a la regresión, denotada SSR.

$$SSR = \sum_{i=1}^{n} (y_i - \bar{y_i})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Un valor grande de SSR indica que la variable independiente sirve para predecir mejor a la variable dependiente.

En regresión logística, el principio es el mismo. En el caso más simple, cuando se tiene solo una variable independiente, se comparan los valores observados de la variable respuesta o dependiente entre dos modelos, siendo el primero el modelo ajustado con la variable independiente -M- y otro sin ella -M_1-. Al igual que cuando se lleva a cabo el ajuste del modelo, en vez de utilizar estadísticas provenientes de un ajuste con mínimos cuadrados, se realizan las pruebas basándose en la log-verosimilitud. Si tomamos

-2 veces el logaritmo de la verosimilitud, denotado -2LL, éste tiene una distribución χ^2. La estadística sin la variable independiente en la ecuación, denotada -2LL inicial, y que corresponde al intercepto del modelo, se denotará D_0, y es análoga a SST. El valor de -2 la log-verosimilitud con intercepto (es decir cuando se toma en cuenta a las variables independientes) es llamado devianza o desviación, y se denota como D. La devianza indica qué tan malo es el modelo cuando se incluye a la variable independiente. La significancia estadística de D es análoga por otro lado a la de la variabilidad no explicada en regresión lineal. Ahora, $G_m = D_0 - D$ corresponde entonces a SSR, y vista como una estadística con distribución χ^2 servirá de prueba para la hipótesis nula $H_0 : \beta_1 = 0$.

Se expresa a G_m como

$$G_m = -2ln\left[\frac{\text{verosimilitud sin la variable}}{\text{verosimilitud con la variable}}\right] \tag{1.9}$$

En el caso de una sola variable independiente, que es el caso que se está trabajando, cuando ésta no está presente en el modelo el estimador máximo verosímil de β_0 es $ln(n_1/n_2)$, donde n_1 es la suma de $y_i's$ y $n_0 = \sum 1/y_i$, además, la predicción del valor de la variable Y es constante: n_1/n.

$$G = -2ln\left[\frac{(\frac{n_1}{n})^{n_1}(\frac{n_0}{n})^{n_0}}{\prod \hat{\pi_i}^{y_i}(1-\hat{\pi_i})^{(1-y_i)}}\right] \tag{1.10}$$

Bajo la hipótesis de que β_1 es cero, para un tamaño de muestra n suficientemente grande, G se distribuye ji-cuadrada con un grado de libertad.

Otro caso especial es cuando el modelo ajustado M es comparado con el modelo saturado o completo M_1. La comparación entre los valores observados y los predichos con la función de verosimilitud se puede ver como:

$$D = -2ln\left[\frac{\text{verosimilitud del modelo ajustado M}}{\text{verosimilitud del modelo saturado } M_1}\right]$$

Esta estadística es el cociente de verosimilitudes, y prueba si los parámetros que están en M_1 pero no en M son iguales a cero. Es decir, suponiendo que para M se tienen $\beta_0, \beta_1, ..., \beta_k$, y para M_1 $\beta_0, \beta_1, ..., \beta_k, \beta_{k+1}, ..., \beta_p$, se tiene la hipótesis nula H_0: $\beta_{k+1}, ..., \beta_p = 0$ contra la hipótesis alternativa H_a: $\beta_i \neq 0$ para $i \in \{k+1, ..., p\}$. Se multiplica por -2 para que se distribuya como una ji-cuadrada. Como ya se mencionó, D es la estadística conocida como devianza, y juega el mismo papel que la suma de cuadrados de los residuales SSE.

En general, es posible comparar entre modelos anidados, siendo el primero el modelo ajustado M_1 y M_0, uno más simple contenido en M_1. La estadística de prueba es el cociente de verosimilitudes, y utiliza las log-verosimilitúdes de ambos modelos. En la sección siguiente (1.4.1) se desarrolla un poco más el concepto del cociente de verosimilitudes.

Regresando al caso en el que el modelo ajustado tiene una sola variable independiente, se puede observar que

$$D = -2ln(\text{verosimilitud del modelo con var. independiente}) \tag{1.11}$$

ya que la verosimilitud de un modelo saturado es

$$l(modelo\ saturado) = \prod_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{(1-y_i)} = 1$$

Ahora, es posible que se reduzca el error en la predicción al usar $\bar{Y}$ a pesar de que no haya relación entre las variables X y Y; ésto debido a la variación de muestreo: fluctuaciones aleatorias en los valores de la muestra que pueden ser engañosas. Para verificar lo anterior, el cociente F sirve para probar si la mejora en la predicción al usar $\hat{Y}$ en lugar de $\bar{Y}$ se puede atribuir a esa variación muestral.

1.4.1. Cociente de verosimilitudes

Supóngase que se quiere probar la hipótesis de que ciertos parámetros son cero. La prueba utiliza la estadística de cociente de verosimilitudes $-2(L_0 - L_1)$ y compara la log-verosimilitud maximizada con dichos parámetros (L_1) con la del modelo simplificado (L_0), en el que dichos parámetros no se toman en cuenta. Denótese por M_1 al modelo ajustado con los parámetros β_i y por M_0 al modelo simplificado. La prueba de bondad de ajuste G^2 es un caso especial del uso de la estadística anterior, en la que el modelo $M_0 = M$, donde M es el modelo ajustado bajo regresión logística, y M_1 es el modelo saturado, es decir el modelo más complejo posible. M_1 es un modelo que tiene un parámetro separado para cada logit y presenta un ajuste perfecto para las logits observadas. A este modelo se le llama el modelo saturado. En este caso la hipótesis nula prueba si todos los parámetros que están en el modelo saturado pero no en el ajustado son cero. A la estadística usada para probar la hipótesis es la devianza, que se puede ver como

$$G^2(M) = L - L_1$$

Para entender al modelo saturado se puede pensar que es un modelo que tiene tantos parámetros β_i's como observaciones. Entonces tenemos la siguiente expresión:

$$D = -2ln\left(\frac{\text{verosimilitud del modelo ajustado}}{\text{verosimilitud del modelo saturado}}\right)$$

Si se ve solamente la expresión dentro del paréntesis, se encuentra al cociente de verosimilitudes entre ambos modelos. Se debe multiplicar por -2 para obtener una estadística que se distribuya aproximadamente como jí-cuadrada y así poder usarla para probar cierta hipótesis. Usando la expresión de la log-verosimilitud, definida como

$$L(\beta) = ln[l(\beta)] = \sum\{y_i ln(\pi(x_i)) + (1 - y_i)ln(1 - \pi(x_i))\}$$

podemos ver a la estadística D como:

$$D = -2\sum(y_i ln(\frac{\hat{\pi}(x_i)}{y_i}) + (1 - y_i)ln(\frac{1 - \hat{\pi}_i}{1 - y_i}))$$

donde $\hat{\pi}_i = \pi(\hat{x}_i)$. Ya se mencionó que la devianza juega el mismo papel en regresión logística que SSE en regresión lineal. De hecho, en el caso en el que la variable dependiente es binaria (toma los valores 1 o 0), la verosimilitud del modelo saturado es 1. Por la definición de éste modelo, se tiene que en este caso $\pi_{ih}at = y_i$, puesto que el modelo saturado proveé un ajuste perfecto. Así, su verosimilitud es

$$l(\text{modelo saturado}) = \prod(y_i^{y_i})(1 - y_i)^{(1-y_i)} = 1$$

De este resultado se sigue que $D = -2ln(\text{verosimilitud del modelo ajustado})$. Para verificar la significancia de una variable independiente X, se compara el valor de D para dos modelos, uno con la variable X y otro sin ésta. El cambio en la devianza D debido a la inclusión de la variable independiente X se obtiene de la siguiente manera:

$$G = D(\text{modelo con X}) - D(\text{modelo sin X})$$

La estadística G se puede ver como

$$G = -2ln\left[\frac{\text{verosimilitud sin X}}{\text{verosimilitud con X}}\right]$$

Ésto se debe a que el modelo saturado se encuentra presente en D(modelo con X) y en D(modelo sin X), que son las dos estadísticas que se están restando. Es por esto que cuando se tiene un ajuste logístico con una sola variable independiente, se compara la devianza de dicho modelo con la de el modelo en el que solo se contiene el intercepto. Cuando la variable no se encuentra en el modelo, el estimado de β_0 es constante y es igual a $ln(n_1/n_0)$ donde $n_1 = \sum y_i$ y $n_0 = \sum(1 - y_i)$. El valor de G será

$$G = -2ln\left[\frac{(\frac{n_1}{n})^n_1(\frac{n_0}{n}{}^n_0)}{\prod \pi_{ih} at_i^y (1 - \pi_{ih} at)^(1 - y_i)}\right]$$

Bajo $H_0 : \beta_1 = 0$, G se distribuye jí cuadrada con un grado de libertad.

1.5. Interpretación del ajuste logístico

Una vez ajustado el modelo y corroborada la significancia de las variables, el enfoque será el de interpretarlo. En esta sección se supone que el modelo ha sido ajustado y que la variable utilizada es estadísticamente significativa. La interpretación debe ser clara para permitir hacer inferencias claras sobre los parámetros estimados en el modelo. La pregunta es que pueden decir entonces los coeficientes sobre la invesigación a la que se debe la modelación. El coeficiente de la variable independiente X representa la pendiente, es decir la tasa de cambio de una función de la variable Y por cambio en una unidad de X. Entonces se debe primero determinar cual es esa unidad de cambio de la variable independiente, y después determinar cual es la relación funcional entre ambas variables. Se procede entonces a determinar que función de la variable dependiente Y, a la que llamaremos función link, arroja una función lineal respecto a X. En el caso de la regresión lineal, la relación ya es lineal, por lo que la función link es la identidad. En regresión logística, la función link es la transformación logit:

$$g(x) = ln(\frac{\pi(x)}{1 - pi(x)}) = \beta_0 + \beta_1 X$$

En regresión lineal, el parámetro β_1 corresponde a la pendiente, y es igual a la diferencia del valor de la variable dependiente Y en $x + 1$ y en $x(\beta_1 = y(x + 1) - y(x))$. La interpretación en este caso es muy directa, pues representa el cambio en las unidades en que medimos Y dado un aumento en una unidad en que se mide X. Por ejemplo, si Y mide altura de un animal en centímetros y X su peso en kilogramos, y la pendiente es 1.5, se concluye que un aumento de un kilogramo en el peso representa un aumento de 1.5 centímetros en su altura.

En regresión logística, la pendiente representa un cambio en el logit correspondiente a

el cambio de una unidad de la variable independiente $X(\beta_1 = g(x+1) - g(x))$. Una correcta interpretación pasará por poder explicar el significado de una diferencia entre dos logits. Primero se ilustra el caso en el que la variable independiente es dicotómica. Éste caso servirá como base para los siguientes. Suponiendo que X está codificada como 1 o 0, la diferencia en los logits para 1 y 0 es $g(1) - g(0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$. La diferencia entre logits es justamente la pendiente. Para poder darle significado a éste y a los resultados que siguen, se introduce el concepto de cociente de momios, llamado *odds ratio* es inglés. Para un individuo tal que x=1, los momios de que la variable respuesta sea un éxito, Y=1, son $\pi(1)/1-\pi(1)$. Análogamente, los momios de un éxito cuando x=0 son $\pi(0)/1 - \pi(0)$. El cociente de momios, denotado OR, está definido como el cociente entre dichos momios:

$$OR = \frac{\pi(1)/1 - \pi(1)}{\pi(0)/1 - \pi(0)}$$

Ahora, $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$, por lo que evaluando en 0 y 1:

- $\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$
- $\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
- $1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$
- $1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$

Por lo tanto, el cociente de momios es $OR = e^{\beta_1}$. OR es una medida de asociación que ha tenido un gran uso especialmente en áreas como epidemología, pues nos dice que tan probablees que la respuesta (codificada por la variable Y) esté presente cuando se observa que en un individuo X=1 comparado con X=0. Para ilustrar esto, supongamos que la variable Y denota la presencia de cáncer de pulmón y X el estátus de fumador del individuo, y obtenemos un OR estimado de 2. En este caso, se puede interpretar el OR como que un fumador tiene dos veces más posibilidades de contraer cáncer de pulmón. La interpretación de OR es así pues aproxima el riesgo relativo, $\pi(1)/\pi(0)$, ya que cuando $\pi(x)$ es pequeño tanto para 1 como para 0,

$$\frac{1 - \pi(0)}{1 - \pi(1)} \to 1.$$

Muchas veces se dicotomiza a una variable continua eligiendo un punto de corte adecuado. Por ejemplo, para CHD y AGE, se puede dicotomizar a la variable explicative AGE como sigue:

AGED= 1 si $AGE \geq 55$ y 0 si $AGE < 55$. La elección de 55 se hace de tal forma que éste valor es la media de AGE.

CUADRO 1.4: Variable AGE dicotomizada según presencia o ausencia de CHD

	AGED(x)		
CHD(y)	$\geq 55(1)$	$< 55(0)$	Total
Presente(1)	21	22	44
Ausente(0)	6	51	57
Total	27	73	100

Codificado así, el modelo logístico arroja los estimados $\hat{\beta}_0 = -0.841$ y $\hat{\beta}_1 = 2.094$. Para éste ejemplo, $\widehat{OR} = e^{2.094} = 8.1$, que se puede interpretar como un aumento de 8.1 en el riesgo de tener CHD para un paciente mayor a 55 años.

1.5.1. Variables continuas

Cuando la variable explicativa es continua, su interpretación depende de como se incluye en el modelo y de su escala de medición. Si se supone que el logit es lineal en la variable llamada X, entonces $g(x) = \beta_0 + \beta_1 x$. El parámetro β_1 da el cambio en los log-momios (o *log-odds*) por cada incremento de una unidad en la variable X. Se puede dar el caso en el que el cambio de una unidad sea muy pequeño o muy grande para interpretarse correctamente, por lo que es importante conocer el problema y la escala de los datos (por ejemplo, si se está modelando la altura en cm contra el peso de un animal, un cambio de un cm puede ser irrelevante, y habría que calcular el cambio para 5 o 10 cm). Se necesita un método para estimar (puntualmente y mediante intervalos de confianza) un cambio arbitrario de c unidades.

$$g(x+c) - g(x) = \beta_0 + \beta_1(x+c) - \beta_0 - \beta_1(x) = \beta_1(c)$$

Obteniendo

$OR(c) = e^{c\beta_1}$

$\widehat{OR} = e^{c\hat{\beta}_1}$

Para estimar el error estándar de $c\beta_1$, se toma

$$e^{c\hat{\beta}_1 \pm z_{1-\alpha/2} c\hat{SE}(\hat{\beta}_1)}$$

1.5.2. Ejemplo (AGE y CHD)

El logit estimado, usando los estimadores ya conocidos, es $\hat{g}(AGE) = -5.31 + 0.111AGE$. Para un aumento de 10 años en la edad del paciente, $\hat{OR} = e^{10(0.111)} = 3.03$. Esto indica que por un incremento de 10 años, el riesgo de contraer CHD aumenta 3.03 veces. Una pregunta que surge es si es lo mismo un cambio de 30 a 40 años que uno de 55 a 65. Si se cuestiona esta interpretación, entonces se está pensando que tal vez la relación de la variable explicativa con el logit no es lineal. En este caso, habría que cambiar el modelo. Si se asume linearidad en el logit, dicha interpretación es inevitable.

Ahora, los intervalos de confianza para $\alpha = 0.05$ son

$$e^{10(0.111)\pm 1.96(10(0.24))} = (1.90, 4.86)$$

1.5.3. Interpretación de valores ajustados y predicciones

Existen casos en el que la interpretación del coeficiente puede ser tan o más importante que la del cociente de momios. En estos casos necesitamos:

- métodos para calcular estimadores e intervalos de confianza
- gráficas para interpretar resultados
- métodos para encontrar predicciónes para individuos fuera de la muestra

Como ejemplo se usan los datos de Hosmer y Lemshaw (2000) del conjunto de datos *Low Birth Weight*, en el que se trata de modelar el nacimiento de un bebé con bajo peso mediante datos de la madre. En este ejemplo la intención es ilustrar el efecto del bajo peso del bebé mediante el peso de la madre usando una raza constante, a saber blanca.

Observense las siguientes gráficas (1.2 y 1.3), que se pueden encontrar en Hosmer y Lemshaw (2000):

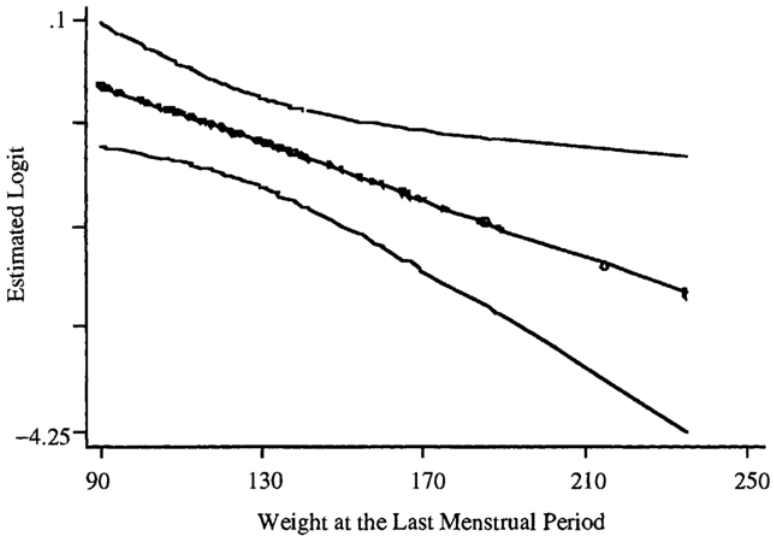


FIGURA 1.2: Logit estimado según el peso

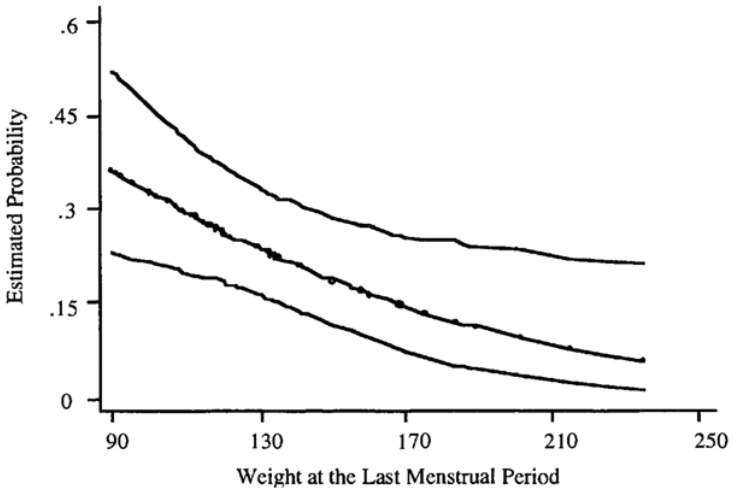


FIGURA 1.3: Probabilidad estimada según el peso

Los estimadores puntuales y por intervalos del logit se pueden transformar en estimadores para la probabilidad logística. Recuérdese que

$$\text{logit: } \hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

y

$$\text{CI: } \hat{g}(x) \pm z_{1-\alpha/2}\hat{SE}(\hat{g}(x))$$

Ahora, para la probabilidad logística,

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

y los intervalos de confianza serán

$$CI = \frac{e^{(\hat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}(\hat{g}(x))}}{1 + e^{\hat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}(\hat{g}(x))}}$$

La probabilidad estimada decrece debido al hecho de que el coeficiente para LBW es negativo Nótese que las bandas de confianza en la gráfica son más angostas entre más cerca se encuentre el valor del peso de la media de LBW (aproximadamente 130 libras). Cada punto, y su intervalo asociado, es un estimado de la media del resultado de Y (bajo peso al nacer) dado el peso X de la madre. Por ejemplo, para una madre de 150 libras, el valor de la media es 0.191 y el intervalo del 95 % de confianza $(0.12, 0.28)$. La probabilidad media de que una madre de ese peso tenga un bebé con bajo peso es de 0.19.

Predicciones

Si se quisiera usar el modelo para estimar la probabilidad de que el peso de un bebé al nacer sea bajo cuando su madre tiene un peso correspondiente de X kilos, de forma que dicho peso no se encuentre en la muestra, se usan los coeficientes estimados para calcular el logit.

Supóngase que la madre es de raza negra y pesa 150 libras, entonces

$$\hat{g}(150, raza = N) = 0.806 - 1.081(150) + 1.081(1) + 0.48(0) = -0.363$$

.

La probabilidad es

$$\hat{\pi}(150, raza = N) = \frac{e^{-0.36}}{1 + e^{-0.36}} = 0.410$$

Dicho resultado se puede interpretar como que el 41 % de las madres negras de 150 libras tendrán bebes con bajo peso de nacimiento (i.e. LBW=1).

1.6. Ejemplo

Para ilustrar una regresión logistica, se modelarán los datos obtenidos del libro *Categorical Data Analysis* (Agresti, 2002). Dichos datos se obtuvieron de un estudio sobre

cangrejos herradura (*horseshoe crabs*). Cada hembra tenía un macho residente en su nido. Se muestran posibles factores que pudieran influir en la presencia de otros cangrejos macho que rodean el nido de una hembra. A éstos se les conoce como "satélites". Las variables son las siguientes:

- color: color del cangrejo (categorizado como 2, 3, 4 o 5)
- spine: condición de la espina dorsal (1=dos en buen estado, 2=una es buen estado,3=ningúna en buen estado)
- width: ancho del cangrejo (cm)
- satellite: número de cangrejos satélite
- weight: peso (gramos)
- Y: presencia (1) o ausencia (0) de satélites

Las siguientes son algunas de las estadísticas descriptivas de los datos obtenidas con R.

```
> summary(crabs)
 color  spine       width        satellite      weight     Y
 2:12   1: 37   Min.   :21.0   0      :62   Min.   :1200   0: 62
 3:95   2: 15   1st Qu.:24.9   3      :19   1st Qu.:2000   1:111
 4:44   3:121   Median :26.1   4      :19   Median :2350
 5:22           Mean   :26.3   1      :16   Mean   :2437
                3rd Qu.:27.7   5      :15   3rd Qu.:2850
                Max.   :33.5   6      :13   Max.   :5200
                               (Other):29
```

Para este ejemplo, se cuenta con observaciones para 173 individuos, y se modelará la existencia de satélites mediante una única variable explicativa, el ancho. Se muestra a continuación una gráfica de los datos. Obsérvese como la variable Y solo toma los valores uno o cero. Esto realmente no nos dá mucha información sobre si es razonable aplicar un ajuste logístico, simplemente se puede concluir que la variable respuesta es binaria.

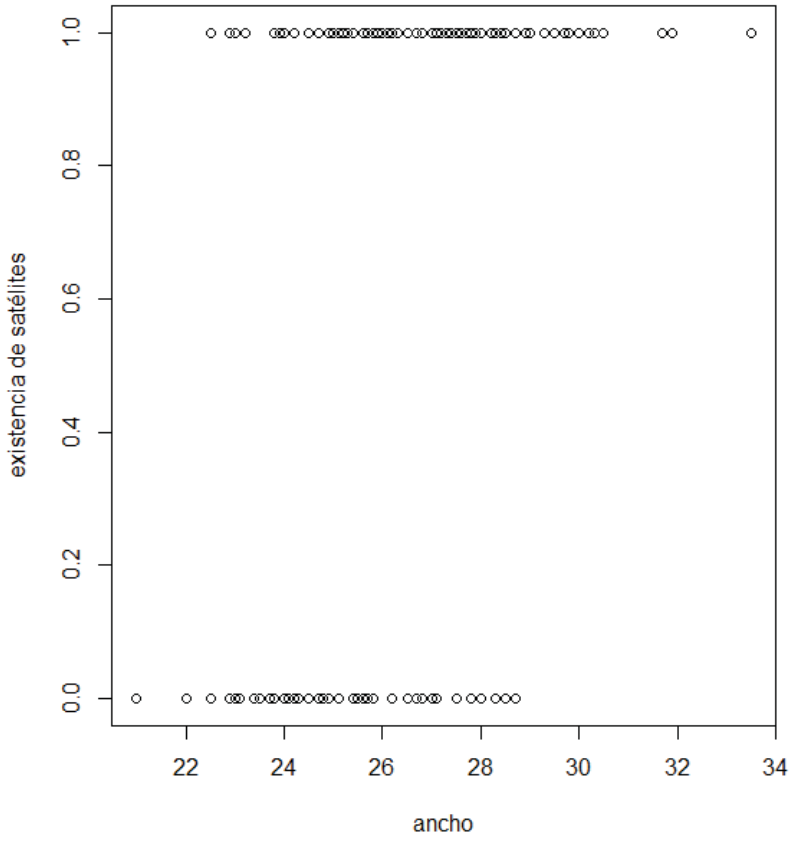


FIGURA 1.4: Gráfica de ancho vs presencia de satélites

Para una variable respuesta binaria Y, y una o varias variables explicativas X, se denota como $\pi(x)$ a la probabilidad de éxito cuando la variable X toma el valor x. Esta probabilidad es entonces el parámetro de una distribución binomial. El modelo de regresión logistico tendrá una forma lineal cuando se toma la transformación logit de dicha probabilidad $\pi(x)$.

$$\text{logit}(\pi(x)) = log\big(\frac{\pi(x)}{1-\pi(x)}\big) = \beta_0 + \beta_1 x \ (1)$$

La fórmula muestra como la probabilidad crece o decrece como una función de x con forma de S, es decir, cuando X decrece, la función se acerca al cero, cuando incrementa, al uno. Otra forma de modelar la regresión logística es usando directamente la exponencial,

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \ (2)$$

El parámetro β_1 determina la tasa de crecimiento de la curva anterior. El signo de β_1 indica entonces si la curva crece o decrece. Cuando el modelo es válido para $\beta_1 = 0$, el lado derecho d la ecuación (2) se reduce a una constante. Así, $\pi(x)$ es idéntica para cada x, convirtiéndose la gráfica de la curva en una recta horizontal. Esto implica que Y es independiente de X. La siguiente gráfica muestra la forma de S de la curva. Dada la forma que toma (y que no se ve una recta), se asume que la función (2) implica que la tasa de cambio en $\pi(x)$ varía en relación a un cambio en X.

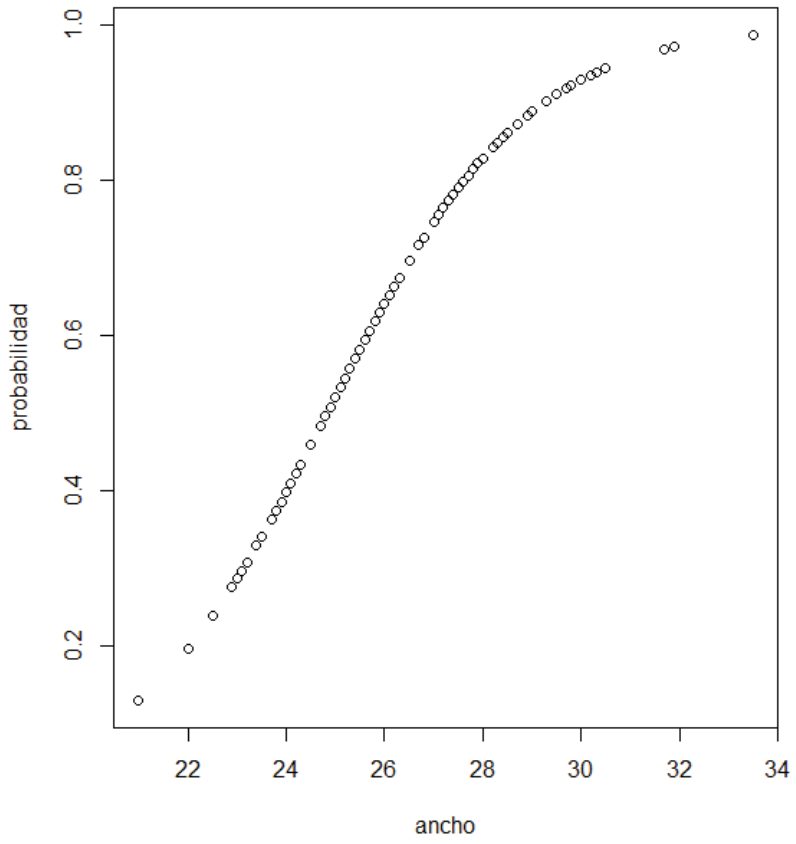


FIGURA 1.5: Gráfica de ancho vs. probabilidad

Sea $\pi(x)$ la probabilidad de que un cangrejo tenga satélites, utilizando el ancho como variable independiente. El modelo más fácil de interpretar es aquel dado por la ecuación (1), pero este modelo tiene problemas. Para estos datos, si se ajusta usando el método de mínimos cuadrados, obtenemos

$$\hat{\pi}(x) = -1.766 + 0.092x$$

La probabilidad según este modelo aumenta en un 0.092 por cada incremento de un centímetro en el ancho. A pesar de que el modelo proveé una interpretación clara y probabilidades realistas para la mayoría de los valores, arroja valores fuera de rango para valores extremos. Por ejemplo, para un ancho de 33.5, la probabilidad estimada de tener satélites sería de 1.3. Por otro lado, cuando se estima a los parámetros utilizando Máxima Verosimilitud, se obtienen valores para los estimadores de $\hat{\beta}_0 = -12.35$ y $\hat{\beta}_1 = 0.497$. La probabilidad predicha utilizando el modelo logístico de la ecuación (2) será, para cada valor de x,

$$\pi(x) = \frac{e^{-12.3508+0.4972x}}{1 + e^{-12.3508+0.4972x}}$$

éste modelo es el que se usa para hacer ajustes bajo regresión logística. A simple vista, el modelo parece ajustarse adecuadamente y es congruente con las observaciones, pero aún así se deben determinar criterios para poder concluir que el ajuste es correcto.

Pruebas de significancia

La devianza indica que tan bueno o malo es el modelo cuando la variable X -que indica el ancho del cangrejo- es incluida en la ecuación. Se calcula también la devianza para el modelo en el que solo hay intercepto (modelo nulo), es decir, para el modelo en el que no se incluye a la variable explicativa. La devianza se calcula como

$$D = -2ln(\text{verosimilitud del modelo})$$

que se distribuye ji-cuadrada con grados de libertad iguales al número de observaciones.

En este ejemplo se obtienen los valores siguientes utilizando el software 'R':

```
    null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
```

Donde la devianza nula corresponde al modelo sin la variable X (ancho) y la devianza residual corresponde al modelo ajustado. La estadística $G = D_0 - D_m$ también se distribuye como jí-cuadrada y sirve para probar la hipótesis de que $\beta_1 = 0$. Si G es tal que el valor-p es menor o igual a cierto valor α, en general 0.05, se rechaza la hipótesis nula H_0 y se concluye que el ancho del cangrejo permite hacer mejores predicciones que si se usara un modelo sin ésta variable.

En este caso se tiene lo siguiente:

$$G = 225.76 - 194.45 = 31.31$$

donde G se distribuye ji-cuadrada con un grado de libertad, resultante de restar los grados de libertad de D_0 (172) y D_m (171). Se tiene un valor-p de 2.204134e-08. El valor-p (llamado también p-value), que representa la probabilidad de que una variable jí-cuadrada con un grado de libertad sea mayor a 31.31, es de hecho es menor que 0.001, por lo que se puede concluir que el modelo ajustado con el ancho explíca mejor la presencia de cangrejos satélite que un modelo nulo, es decir un modelo sin variables explicativas.
Lo anterior es evidencia estadística de la significáncia del ancho como predictor de la presencia de satélites, pero antes de sacar conclusiones definitivas se debe verificar que el ajuste del modelo sea bueno.

Prueba de Wald

Existe una prueba estadísticamente equivalente a la del cociente de verosimilitudes, la prueba de Wald, que compara el estimador máximo verosímil de la pendiente β_1 (el coeficiente de la variable independiente o explicativa, en este caso el ancho), con un estimado de su error estándar. Bajo la hipótesis de que $\beta_1 = 0$, la estadística se distribuye normal estándar.

$$W = \frac{\hat{\beta}_1}{\widehat{SE}\hat{\beta}_1}$$

Para éste ejemplo, el estimado es $\hat{\beta}_1 = 0.4972$ con error estándar $\widehat{SE} = 0.1017$, por lo que

$$z = \frac{0.4972}{0.1017} = 4.888889$$

El valor-p de dos colas representa $P(|z| > 4.88)$, donde z sigue una distribución normal estándar. En este caso se obtiene un valor-p de 0.0072, por lo que a un nivel de significancia de 0.05 se rechaza la hipótesis nula.

Se continúa por interpretar los resultados de los parámetros estimados. Recuérdese que el logit es el siguiente

$$g(x) = \beta_0 + \beta_1 x.$$

El parámetro β_1 da el cambio en los log-momios, que podemos ver como

$$OR(c) = e^{c\hat{\beta}_1}$$

para un valor arbitrario de c. El valor de c depende de la escala de la variable independiente y de como interpretemos el modelo. En este caso, dado que los datos varían entre 23 y 30 centímetros aproximadamente, parece correcto considerar un cambio en un centímetro. Ajustando el valor estimado de β_0, y fijando el valor de c como 1 centímetro, el cociente de momios es el siguiente: $\widehat{OR(1)} = e^{0.4972} = 1.644111$, donde $\hat{\beta}_1 = 0.4972$. Los extremos de un intervalo del 95 % de confianza, donde se usa un valor de $\alpha = 0.05$ son

$$e^{c\hat{\beta}_1 \pm z_{1-\alpha/2} c\hat{SE}(\hat{\beta}_1)}$$

Evaluando utilizando los estimadores se obtiene lo siguiente:

$$(e^{0.4972-1.960.1017}, e^{0.4972+1.960.1017} = (1.346984, 2.006781)$$

Se interpreta a los momios como sigue: para un aumento de un centímetro en el ancho del cangrejo, la probabilidad de tener satélites aumenta en 0.016331. Por ejemplo, es 1.64 veces más probable que un cangrejo de 25 centímetros tenga satélites comparado con uno de 24.

1.7. Otras consideraciones

A pesar de que el capítulo trata principalmente la regresión logística para una variable respuesta binaria, y en gran parte se desarrolla el modelo utilizando una sola variable explicativa, no sobra mencionar que también es posible generalizar la regresión logística para variables respuesta multinomiales. En el caso en que $Y = i$ con $i \in 1, ..., J$, se busca modelar la probabilidad de que Y sea igual a cierta categoría, dada una observación $x = (x_1, ..., x_k)$:

$$\pi_j(x) = P(Y = j|X = x)$$

donde $\sum_{j}^{J} \pi_j(x) = 1$. En este caso se supone que Y es una variable multinomial, con probabilidades $\pi_1(x), ..., \pi_J(x)$. Suponiendo que se toma a la categoría J como base, se procede a modelar de la siguiente manera,

$$ln\frac{\pi_j(x)}{\pi_J(x)} = \beta_0 + \beta_j' x$$

Obsérvese que con estas $J-1$ ecuaciones es posible comparar entre cualquier par de categorías a y b, ya que

$$ln\frac{\pi_a(x)}{\pi_b(x)} = ln\frac{\pi_a(x)}{\pi_J(x)} - ln\frac{\pi_b(x)}{\pi_J(x)}$$

Ahora, para estimar las probabilidades de respuesta,

$$\pi_j(x) = \frac{e^{(\alpha_j+\beta_j' x)}}{1+\sum_{h=1}^{J-1} e^{(\alpha_h+\beta_h' x)}}$$

donde $\alpha_J = 0$ y $\beta_J = 0$. Se utiliza máxima verosimilitud para $J-1$ ecuaciónes que se optimizan numéricamente, o que se estiman a través de ajustar por separado logits binarios para $J-1$ parejas de respuestas.

En el caso en el que los valores de las categorías de Y son ordinales, es decir están ordenados y se presentan tendencias monotonas crecientes o decrecientes, la manera más común de modelar es mediante el uso de logits acumulados (*Cumulative Logits*). La idea es utilizr logits de probabilidades acumuladas,

$$P(Y \leq j|x) = \pi_1(x) + ... + \pi_j(x)$$

para $j \in 1, ..., J$ y donde se define a los logits acumulados como sigue:

$$logit[P(Y \leq j|x)] = ln\frac{P(Y \leq j|x)}{1-P(Y \leq j|x)}$$
$$= ln\frac{\pi_1(x) + ... + \pi_j(x)}{\pi_{j+1}(x) + ... + \pi_J(x)}$$

.

Nótese que un modelo como éste, pero en el que sólamente se busca el $logit[P(Y \leq j|x)]$ se puede ver como un modelo logit ordinario para una respuesta binaria, donde las categorías 1,...,j forman un resultdo, y j+1,...,J el otro. De hecho es posible obtener un modelo que usa simultaneamente los J-1 logits acumulados. Éste se ve como $logitP(Y \leq j|x)=$

$\alpha_j + \beta_j' x$ para $j \in 1, ..., J-1$. Cada logit acumulado tendrá su propio intercepto y éstos serán crecientes en j para una x fija. Los efectos β serán iguales.

Así como éstos dos modelos, que simplemente fueron ilustrados a manera de ejemplo, existen otras variantes del modelo logístico o de los modelos lineales generalizados para variables categóricas y estratégias para su construcción. Para profundizar más en ellos, tanto en la estimación de los parámetros, las pruebas de bondad de ajuste, aplicaciones, u otros temas más específicos, véase Agresti, 2002, cáp. 7 ó Hosmer y Lemshaw, 2000, cáp. 8.

Capítulo 2

Árboles de clasificación y regresión

Los métodos desarrollados a partir del uso de árboles pueden ser usados para analizar datos cuando el objetivo es clasificarlos o determinar una regresión. En el caso de la regresión logística, se puede ver el objetivo desde los dos puntos de vista: primero se requiere clasificar a la variable respuesta dentro de una de las k categorías (este trabajo se enfoca en regresión logística binaria), pero no se agota el problema clasificando, pues parte importante es también identificar la relación entre las variables independientes y la variable respuesta. El algoritmo que se estudiará en éste segundo capítulo -desarrollado por Breiman, Freidman, Olshen y Stone en 1984- es el algorítmo CART *(Clasification And Regression Trees)*, en el que se desarrolla un método para generar un árbol de decisión a partir de un conjunto de aprendizaje. Dicho árbol será utilizado después para clasificar nuevos datos, o en el caso de la regresión, para predecir el valor esperado de la variable respuesta.

2.1. Clasificación

El problema de clasificar consiste básicamente en utilizar las mediciones o datos observados de un conjunto de variables, y utilizarlas para predecir a que clase pertenece un objeto. Entonces, dado un conjunto de variables correspondientes a una serie de observaciones, el objetivo es dar un método sistemático de predecir a que clase pertenece cada observación.

Los árboles binarios clasificadores son construidos mediante divisiones del conjunto de datos. Si se supone que se tiene una muestra aleatoria de n observaciones $(y_i, x_1 i, ..., x_k i)$,

el conjunto inicial es el total de las observaciones, que se particionará mediante divisiones implementadas progresivamente.

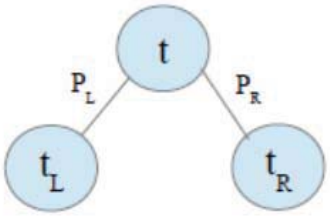


FIGURA 2.1: Ejemplo 1

Por ejemplo, si se llama al conjunto de todas nuestras observaciones X, la división lo particionará en X_1 y X_2 disjuntos, donde $X = X_1 U X_2$. Las divisiones son formadas mediante la imposición de condiciones en las variables independientes. Por ejemplo se podría formular la siguiente división:

Se denota al conjunto de variables independientes como $X = (x_1, ..., x_k)$, entonces se busca el conjunto $\{x_2 \leq 7\}$ Las observaciones que cumplan la condición se irán al nodo t_1 mientras que las que no a t_2.

$t_1 = \{x | x_2 \leq 7\}$
$t_2 = \{x | x_2 > 7\}$

Así, se continúa ahora dividiendo a t_1 y t_2, hasta que se llega a un nodo terminal (más tarde se verá como se determina cuando un nodo es terminal). La clase predicha para las observaciones correspondientes a cada nodo terminal será determinada también por una clase previamente elegida para cada nodo.

Se presenta a continuación la terminología que se usará a partir de este momento, y que corresponde a aquella dada por la teoría de árboles:

- un nodo t = un subconjunto de X
- el nodo raíz t_1 = X.
- nodos terminales = subconjuntos terminales

A gran escala se puede considerar que la construcción de un árbol, ya sea de clasificación o de regresión, se determina mediante los siguientes pasos:

1. selección de las divisiones

2. decisión para determinar cuando un nodo es terminal

3. asignación de una clase o categoría a cada nodo terminal

El conjunto de datos será utilizado tanto para determinar las divisiones, los nodos terminales y sus asignaciones correspondientes.

2.2. Metodología

Para un problema de clasificación en el que tenemos J clases distintas, llámese $\mathcal{L}$ a la muestra de aprendizaje de tamaño n; es decir, $\mathcal{L}$ es la muestra que se usará para construir el árbol. Sea entonces n_j el número de elementos en la clase j para $j = 1, ..., J$. Las probabilidades a priori $\pi(j)$, a menos que sean ya proporcionadas, suelen tomarse de acuerdo a las proporciones n_j/n, aunque éstas no necesariamente reflejan las proporciones esperadas a futuro. En un nodo cualquiera t, sea n(t) el total de elementos de $\mathcal{L}$ en t y $n_j(t)$ el número de dichos elementos pertenecientes a una clase j. Así, la proporción de los elementos de la clase j será análogamente $n_j(t)/n(t)$. Las probabilidades a priori $\pi(j)$ se pueden interpretar como la probabilidad de que una observación nueva pertenezca a la clase j, por lo que el estimado de resustitución, definido como la probabilidad de que un elemento pertenezca a t y a la clase j, es $p(j,t) = \pi(j)n_j(t)/n(t)$. Entonces el estimado de resustitución $p(t)$ de la probabilidad de que cualquier caso pertenezca al nodo t es

$$p(t) = \sum p(j,t) \tag{2.1}$$

El estimado de resustitución de la probabilidad de que un elemento esté en la clase j dado que está en el nodo t es

$$p(j|t) = p(j,t)p(t) \tag{2.2}$$

y se satisface que $\sum p(j|t) = 1$. Cuando $\pi(j) = n_j/n$ es dada, entonces $p(j|t) = n_j(t)/n(t)$, por lo que $p(j|t)$ son las proporciones relativas de la clase j dentro del nodo t. Recuerdese que para iniciar el proceso de construcción del árbol se necesita lo siguiente:

- un conjunto de preguntas Q de la forma $\{x \in A|A \subset X\}$.
- un criterio de bondad de división $\phi(s,t)$ donde s es una división y t un nodo cualquiera.

- una regla para detener la división.
- una regla para asignar una clase a cada nodo terminal.

2.2.1. Preguntas

El conjunto de preguntas Q se asigna de la siguiente forma. Suponiendo que los vectores de variables independientes tienen la forma

$$x = (x_1, ..., x_M)$$

donde M se puede ver como el número de variables independientes o como la dimensión (fija) del vector x, se define a un conjunto estándar de preguntas Q como sigue:

1. Cada división depende únicamente del valor de una sola variable x_i con

$$i \in \{1, ..., M\}.$$

2. Para cada variable ordenada x_i, Q incluye todas las preguntas de la forma $\{x_i \leq c\}$ donde $c \in (-\infty, \infty)$.

3. Para cada variable categórica x_i que toma valores en un conjunto $A = \{A_1, ..., A_L\}$, Q incluye todas las preguntas $\{x \in A_i | A_i \subset A\}$.

Nótese que la cantidad de divisiones posibles es finita, pues para cada variable x_i el conjunto de observaciones contiene a lo más n distintos valores para x_i, por lo que hay a lo más n divisiones distintas generadas por las preguntas de Q sobre x_i, a saber las de la forma $x_i \leq c_j$ con $j = 1, ..., n$, para los valores c_j tomados de tal forma que $x_i < c_j < x_{i+1}$ con $i = 1, ..., n-1$. Se puede ver que análogamente para variables categóricas se tiene a lo más $2^{L-1} - 1$ divisiones posibles, donde L es el número de categorías. El algorítmo encuentra para cada variable x_i la mejor división, y después compara entre las variables, eligiendo así en cada nodo la mejor división s.

Como ejemplo considérese un problema de dos clases en el que hay dos variables independientes x_1 y x_2 con $0 < x_i < 1$ para $i = 1, 2$. Si se supone que el árbol se ve como en la figura 2.1, una forma geométrica de interpretarlo es considerarlo como una partición del cuadrado unitario. Desde éste punto de vista, se puede interpretar como una partición recursiva del espacio, con la cual se busca homogeneizar a las poblaciones dentro de cada partición.

2.2.2. Divisiones

El primer problema es elegir adecuadamente las divisiones de X en subconjuntos cada vez más pequeños. La idea es elegir de tal forma que los datos contenidos en cada uno de los nodos descendientes sean más puros que los datos del nodo padre. Esta idea de pureza corresponde al número de clases que se encuentran en un nodo; por ejemplo, si contamos con cuatro clases mezcladas dentro de un nodo t, sería ideal que la división asignara a todos los elementos pertenecientes a las primeras dos clases al nodo t_1 y a todos los demás a t_2. En la figura 2.2 se ilustra la idea anterior.

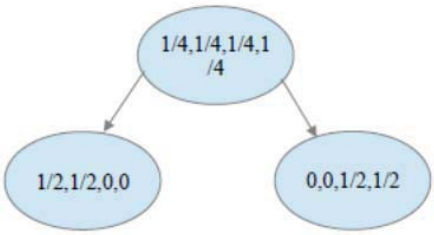


FIGURA 2.2: Ejemplo 2

Una posible implementación es la siguiente:

1. Definir las proporciones nodales en el nodo t $p(j|t)$ como la proporción de los casos $x_n \in t$ pertenecientes a la clase j, donde se supone que $j = 1, ..., J$, tales que

$$p(1|t) + ... + p(J|t) = 1$$

2. Definir una medida de impureza del nodo t como una función no negativa de las $p(j|t)$,

$$\phi : \{p(j|t)\} \rightarrow [0, \infty)$$

La función ϕ debe cumplir que la impureza sea máxima cuando todas las clases estén mezcladas homogeneamente, y debe alcanzar su mínimo (a saber 0) cuando hay solo una clase perteneciente al nodo t.

Para determinar las divisiones o la partición de un nodo t en t_L y t_R se ha definido el decrecimiento en impureza como

$$\Delta i(s, t) = i(t) - (P_L i(t_L) + P_R i(t_R)).$$

En este caso utilizaremos la siguiente forma para la función de impureza de un nodo t:

$$i(t) = \sum \phi(p_{it}) \tag{2.3}$$

donde p_{it} es la proporción de casos pertenecientes a la clase i dentro del nodo t y la función ϕ tiene dos posibles formas: el índice de información, $\phi(p) = -pln(p)$ o el índice Gini, definido como $\phi(p) = p(1-p)$. La justificación de la forma de dichas funciones se deriva de las propiedades que debe cumplir una función de impureza vistas previamente. Se elige entonces para cada nodo la variable que genere la mayor reducción en impureza.

El modelo CART incorpora dentro del criterio de reducción de riesgo la función de pérdida. Ésto se hace mediante la implementación de las llamadas probabilidades a priori modificadas. Si $L(i,j)$ es una medida del costo del error al clasificar a una observación de la clase i como perteneciente a la clase j, y el riesgo de t es:

$$R(t) = \sum p(i|t)L(i,\tau\,(t)) \tag{2.4}$$

donde $\tau(t)$ -la clase asignada a t- se elige para minimizar $R(t)$. Entonces:

$$R(t) = \sum \pi_i L(i,\tau\,(t)) \frac{n_i t}{n_i} \frac{n}{n_t} \tag{2.5}$$

Supongase que existen L^* y $\hat{\pi}_i$ tales que $\hat{\pi}_i L^*(i,\tau\,(t)) = \pi_i L(i,\tau\,(t))$. Entonces la función de riesgo R no se vería alterada. Si L^* es proporciónal a la matríz de pérdida de ceros y unos, entonces se utilizan las probabilidades a priori $\hat{\pi}_i$ como criterio para dividir a un nodo, en cuyo caso

$$\hat{\pi_i} = \frac{\pi_i L_i}{\sum\limits_j \pi_j L_j}$$

La elección de probabilidades a priori solo altera la elección de la división del nodo, eligiendo mejores divisiones en términos del riesgo.

El paso final consiste en definir un conjunto de divisiones S en cada nodo. Se puede pensar que el conjunto S se genera mediante un conjunto correspondiente de preguntas o condiciones Q, donde Q genera una partición de la forma

$\{x \in A | A \subset X\}$ o $\{x \notin A | A \subset X\}$

Entonces, la división s asociada a dicha pregunta manda a todas las x_n que satisfacen la condición al nodo t_l y al resto a t_r. Para determinar cuando detener el crecimiento del árbol se suele determinar alguna regla heurística. Por ejemplo, al alcanzar un nodo en el que se considera no significante el decrecimiento en la impureza, ese nodo no se divide y se considera un nodo terminal. Después, se prosigue a caracterizar el nodo terminal con alguna de las clases $j = 1, ..., J$. Un buen ejemplo de como se puede hacer ésto es asignar a aquella la clase j tal que el número de elementos pertenecientes al nodo terminal correspondientes a j sea máximo.

Anteriormente se definió la medida de impureza en base a una función ϕ. El concepto de bondad de la división se deriva de ésta función ϕ, a la que se llama función de impureza. La función de impureza actúa sobre un vector de la forma $(p_1, ..., p_J)$ tal que $p_i \geq 0$, y $\sum p_i = 1$, y que cumpla las propiedades:

- ϕ alcanza su máximo solo en el punto $(1/j, 1/j, ..., 1/j)$
- ϕ alcanza su mínimo en los puntos $(1, 0, ..., 0)$, $(0, 1, ..., 0)$, $(0, 0, ..., 1)$.
- ϕ es simétrica para $p_1, ..., p_J$

Ahora, para todo nodo t, la medida de impureza $i(t)$ fué definida como:

$$i(t) = \phi(p(1|t), ..., p(J|t))$$

Si una división s del nodo t manda una proporción P_R de los datos en t a t_R y a una proporción P_L de los datos en t a t_L, el decrecimiento en impureza será

$$\Delta i(s, t) = i(t) - (P_L i(t_L) + P_R i(t_R))$$

y se toma aΔ $i(s, t)$ como la bondad de la división $\psi(s, t)$.

Ahora, si el conjunto de nodos terminales de un árbol T es denotado como $\tilde{T}$, donde el árbol T está determinado por el conjunto de divisiones s utilizadas y por el orden de éstas, sea $I(t) = i(t)p(t)$. Se define la impureza del árbol T como

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)p(t) \tag{2.6}$$

La elección de las divisiones s que maximizan a$\Delta\, i(s,t)$ equivale a la elección de aquellas que minimizan la impureza general del árbol $I(T)$. Considerese entonces a la selección de las divisiones como intentos por ir minimizando la impureza. Para detener las divisiones se debe determinar una regla heurística. En este caso por ejemplo se declara un número $\beta > 0$ y se dice que un nodo t es terminal cuando

$$max_{s\in S}\Delta i(s,t) < \beta$$

2.2.3. Asignación de clases

Supongase que se construye el árbol T, y que el conjunto de sus nodos terminales es $\tilde{T}$. Una regla de asignación de clase asocia una de las clases $j \in 1, ..., J$ a cada uno de los nodos terminales $t \in \tilde{T}$. A la clase asociada a un nodo t se le denota j(t). Cuando las probabilidades a priori $\pi(j) = n_j/n$, se clasifica a t con la clase para la que $n_j(t)$ sea mayor. Para cualquier conjunto de probabilidades a priori $\pi(j)$ y una regla de asignación de clase $j(t)$, el valor

$$\sum_{j\neq j(t)} p(j|t)$$

es llamado el estimado de resustitución de la probabilidad de un error al clasificar dado que una observación se encuentra en el nodo t. La regla de asignación de clase $j^*(t)$ es aquella que minimiza dicho estimado de resustitución. Es decir, la regla de asignación de clase $j^*(t)$ esta dada por la siguiente condición:

- Si $p(j|t) = max_i p(i|t)$, entonces $j^*(t) = j$. Si el máximo se alcanza por dos o más clases, se determina arbitrariamente cuál de éstas es la clase del nodo. Así, se obtiene que el estimado de resustitución $r(t)$ de la probabilidad de clasificar mal dado que la observación pertenece al nodo t es

$$r(t) = 1 - max_j p(j|t)$$

Si se denota como $R(t) = r(t)p(t)$, entonces el estimado general de la tasa de error de clasificación $R^*(T)$ para el árbol T es $R^*(t) = \sum_{t\in\tilde{T}} R(T)$.
Una propiedad importante de $R(T)$ es que irá decreciendo entre más divisiones tengamos. Es decir, si el árbol T' se obtiene a partir de un árbol T al que se le añade un nodo tras dividir algún nodo terminal de T, entonces $R(T') \leq R(T)$. Visto de otra forma, si se divide a un nodo t en t_L y t_R, entonces $R(t) \geq R(t_L) + R(t_R)$

2.2.4. Poda y estimación

El proceso de construcción de los árboles termina cuando se obtiene un árbol T_{max} con el número máximo posible de nodos. Se busca que todos los nodos terminales sean puros, es decir que solo contengan observaciones de la misma categoría, o que su tamaño sea lo suficientemente pequeño (se puede especificar un número N_{min} y determinar que un nodo t es terminal cuando $N_{(t)} \leq N_{min}$, donde $N_{(t)}$ es el número de observaciones en el nodo t).

Ahora, se llama a un nodo t' descendiente de t si existe un camino que conecte a t' con t. La rama T_t de un árbol T se forma con el nodo t y sus descendientes. Se dice que podamos la rama T_t cuando a todos los descendientes del nodo t son eliminados.

Si el árbol T' se obtiene de podar a T, es decir de quitarle alguna rama, se le llama a T' un subárbol de T, y se denota como $T' < T$.
Ahora, $R(T)$, el estimado de resustitución para el costo de error de clasificación $R^*(t)$ es

$$R(T) = \sum_{t \in \tilde{T}} R(t) = \sum_{t \in \tilde{T}} r(t)p(t) \tag{2.7}$$

$R(T)$ es un buen criterio para comparar entre árboles o subárboles del mismo tamaño. Al ir podando, se debe ir obteniendo una secuencia de subárboles tales que en cada etapa de la poda obtengamos al de menor estimado de resustitución. Es decir, si se genera la secuencia $T_{max} > T_2 > ... > t$, y $n_i = |\tilde{T}_i|$ es el número de nodos terminales del subárbol T_i, se busca que para todo árbol T_j con $n_j = n_i$, $R(T_i) \leq R(T_j)$, es decir, T_i debe ser el subárbol de mínimo costo con n_i nodos terminales. El problema de generar a la secuencia de subárboles de ésta forma es que la secuencia no es necesariamente anidada: si se cuenta con la secuencia $T_i > T_{i+1}$, el árbol T_{i+1} no tiene por que haberse obtenido de podar a T_i.

2.2.5. Poda de mínimo costo-complejidad

Para solucionar el problema anterior, se lleva a cabo el proceso de poda de mínimo costo-complejidad. Si se define la complejidad de un árbol T como $|\tilde{T}|$, el número de nodos terminales, y el parámetro de complejidad $\alpha \geq 0$, entonces el estimado de costo-complejidad es

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{2.8}$$

La idea es añadir a $R(T)$ un número que pondere la complejidad de T. Para cada valor de α, se busca el subárbol que minimíce $R_\alpha(T)$. Dicho subárbol se denotará como $T(\alpha)$, y R_α es aquel valor tal que $R_\alpha(T(\alpha)) = min_{T<T_{max}} R_\alpha(T)$.

Una vez que se tiene un árbol máximo, se procede a decidir que tanto de ese árbol retener, y eliminar las particiones que no son necesarias: a saber las que no aportan información relevante o solo complican la interpretación. Supóngase que se tiene un conjunto de nodos terminales $\tilde{T}$. El parámetro de complejidad α es una medida del costo de añadir otro nodo al modelo. Si $R(T_0)$ es el riesgo de un árbol sin divisiones (es decir un árbol con solo el nodo raíz), entonces el costo del árbol T se definió como

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

T_α es definido como aquél subárbol con mínimo costo para el parámetro α. Se sigue de la definición de α que $T_{\alpha=0}$ =modelo completo y
$T_{\alpha\to\infty}$ = modelo sin divisiones. Además:

- Si T_1 y T_2 son subárboles de T con $R_\alpha(T_1) = R_\alpha(T_2)$ entonces T_1 es subárbol de T_2 o viceversa.
- Si $\alpha > \beta$ entonces $T_\alpha = T_\beta$ o T_α es subárbol de T_β

Se puede entonces definir únicamente al subárbol T_α como el menor árbol que minimiza a $R_\alpha(T)$. Además es posible agrupar en m intervalos con $m \leq |\tilde{T}|$ a los posibles valores de α, debido a que cualquier secuencia de subárboles del árbol T tiene a lo más $|\tilde{T}|$ nodos terminales. Se denota como $I_1, ..., I_m$ a dichos intervalos, donde para toda $\alpha \in I_j$, α contiene al mismo subárbol que minimiza $R_\alpha(T)$. Para elegir el mejor valor de α, usamos el método de validación cruzada. Se calculan primero los intervalos I_j para el árbol máximo, y a un α_j correspondiente. Para cada α_j se encuentra mediante el método de validación cruzada un estimado del riesgo, y se selecciona a β^* como aquel parámetro que arroje el menor valor del estimado de riesgo. El mejor árbol podado será entonces T_α, tomando el modelo completo T_{max}.

Nótese que si el valor de α es pequeño, la penalización por tener muchos nodos terminales también lo será, pero el subárbol T_α será grande. Así, si se genera al árbol máximo T_{max} de manera que todos los nodos terminales tengan solo una observación, entonces $R(T_{max}) = 0$, pues toda observación es clasificada correctamente, pero T_{max} tendrá el mayor número posible de nodos terminales. Al ir aumentando el valor de α, los subárboles que minimizan $R_\alpha(T)$ van a ir teniendo cada vez menos nodos terminales. Por ejemplo,

para $\alpha = 0$, $R_0(T(0)) = 0$ si y solo si $T(0) = T_{max}$, y entonces se eligiría a T_{max} como mejor subárbol. Por otro lado, para un valor de α suficientemente grande, $T(\alpha)$ será $\{t\}$, el subárbol con solo el nodo raíz.
A pesar de que α puede tomar valores en R, existe solo un número finito de subárboles pues el número de nodos terminales máximo es finito, y para cada valor posible de nodos terminales solo hay un mejor subárbol. Si $T(\alpha)$ es el mejor subárbol para un α dado, seguirá siéndolo hasta llegar a un punto de quiebre α'. El algorítmo para encontrar el mejor subárbol es descrito con detalle en Breiman (cap. 3.3 y 3.4).

2.2.6. Elección de probabilidades iniciales y costos de error de clasificación variables

Para una muestra de aprendizaje inicial L con J clases, sea N_j el número de observaciones pertenecientes a la clase j $\in 1, ..., J$. Usualmente las probabilidades iniciales o a priori $\pi(j)$, que representan la probabilidad inicial de que una clase esté presente en el árbol, son tomadas como las proporciones muestrales N_j/N. Recuerdese que el estimado de resustitución, que representa la probabilidad de que una observación pertenezca a la clase j y a un nodo t es el siguiente:

$$p(j,t) = \pi(j)(N_j(t)/N_J)$$

El criterio para separar o particionar a un nodo t se basa en la medida de impureza i(t). De hecho la construcción de un árbol se basa en la idea de generar árboles con la menor impureza posible. La medida de impureza para un nodo t,

$$i(t) = \phi(p(1|t), ..., p(J|t))$$

depende directamente de las probabilidades iniciales elegidas, ya que $p(i|t)$ depende directamente de la probabilidad inicial $\pi(i)$.
Por otro lado, se tiene también que el estimado de resstitución $r(t) = 1 - max_{j=1}p(j|t)$ y $R(t) = r(t)p(t)$. Es claro que éste también depende de la elección de las probabilidades iniciales $\pi(j)$. Una buena elección de probabilidades a priori, siendo parámetros que pueden ser seleccionados antes de comenzar la construcci'on del árbol, puede asistir a que éste tenga una estructura final adecuada. A pesar de que se suelen elegir las proporciones muestrales, hay casos en los que los datos pueden no estar balanceados entre clases, lo que produce una gran disparidad entre el costo de clasificar mal a cada

clase, lo que a su vez produce un resultado final equivocado. Por ejemplo, en los datos encontrados en el estudio de espectro de masa (véase Breiman, p. 203), la proporción entre compuestos con y sin cloro es de 10:1. En este ejemplo, si se utilizan las proporciones muestrales como probabilidades iniciales, se obtiene un árbol donde todo compuesto se clasifica como si no contuviera cloro. ejemplo ilustra como las probabilidades a priori pueden ser elegidas de forma que se ajuste el costo de error de clasificación entre clases.

Costos de error variables

En el caso de la medida de impureza, que se utiliza para el criterio de división, es posible utilizar una extensión de el índice de Gini como función de impureza. La idea es que es posible especificar costos de error de clasificación variables $C(i|j)$ de forma que el costo de clasificar a una observación perteneciente a la clase i como perteneciente a la clase j sea distinta para cada j $\in 1, ..., J$ (para más detalles, véase Breiman, 4.4.2).

La construcción final de un árbol de clasificación dependen de la elección de las divisiones y de la poda. Como es posible observar, éstas dos (a través de la medida de impureza y el costo de error de clasificación) dependen fuertemente de la elección de las probabilidades iniciales $\pi(j)$, ya que pueden utilizarse para ajustar el resultado final del árbol. Es por ésto que es importante considerar con cuidado si éstas se deben tomar como las proporciones muestrales o como alguna versión modificada. En todo caso, es importante darse cuenta que se debe tener un conocimiento previo de los datos con los que se está trabajando y objetivos claros de lo que se busca para lograr resultados adecuados.

2.3. Árboles de regresón

En regresión, una observación i consiste de los datos (x_i, y_i) donde x_i es un vector de la forma $x_i = (x_i1, ..., x_iM)$. La variable Y es la variable respuesta o variable dependiente, mientras que las variables en X son las variables independientes o explicativas. El objetivo de la regresión es construir una función $d(x)$ para entender las relaciones estructurales entre variables independientes y la variable dependiente, así como predecir valores futuros correspondientes a la variable y, dados los valores de x. La diferencia principal con el problema de clasificar variables es que en regresión se trata de determinar la naturaleza de la relación de las variables observadas, relación que será determinada por el predictor $d(x)$. Los casos más comunes de regresión son la regresión lineal y la logística. Primero se desarrolla el método general para construir árboles de regresión, para después ver como se pueden modificar para ser implementados en regresión logística binaria.

Considérese una muestra de aprendizaje $\mathcal{L}$ con n observaciones $(x_1, y_1), ..., (x_n, y_n)$ con la que se va a construir el predictor $d(x)$. Para entender la función del predictor se puede usar un ejemplo de regresión lineal, donde se asume que

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

En este caso, el predictor sería $d(x) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$. Llámese entonces $\hat{y}$ al predictor de y, para continuar con la notación del capítulo previo. Se puede entonces medir la eficiencia del predictor mediante el error promedio,

$$\frac{1}{n}\sum(y_i - d(x_i))^2 = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

Esta medida no es más que la suma de cuadrados de los errores, llamada SSR. La construcción de árboles de regresión basándose en ésta medida es llamada regresión por mínimos cuadrados.

Se define ahora al error cuadrático medio $R^*(d)$ del predictor d como,

$$R^*(d) = E(Y - d(X))^2 \tag{2.9}$$

Es decir, $R^*(d)$ es el error esperado al cuadrado cuando usamos a $d(X)$ como predictor de Y. En el caso de los árboles de clasificación, $R^*(d)$ denota la tasa de error de clasificación. Se usa la misma notación, pues ambas son medidas de eficiencia de la predicción, ya sea de la variable respuesta o de la clase asociada.

El predictor $d_B(x)$ que minimiza a $R^*(d)$ es $d_B(x) = E(Y|X = x)$, la esperanza condicional de la respuesta Y dados los valores del vector x. Por otro lado, la constante que minimiza a $E(Y - a)^2$ es $E(Y)$.

2.3.1. Medidas de error y sus estimados

Dada una muestra de aprendizaje $\mathcal{L}$ de la forma $(x_1, y_1), ..., (x_n, y_n)$, se quiere utilizarla tanto para construir el predictor como para evaluarlo estimando su error $R^*(d)$. Hay muchas maneras de hacer lo anterior, aquí se mostraran las tres más comunes. La primera es el estimado de resustitución,

$$R(d) = \frac{1}{n}\sum(y_i - d(x_i))^2 \tag{2.10}$$

Otra forma de evaluar es dividiendo la muestra $\mathcal{L}$ en una muestra de aprendizaje $\mathcal{L}_1$ de tamao n_1 con la que se construye el predictor $d(x)$ y una de prueba $\mathcal{L}_2$ para evaluarlo, de tamaño n_2, para construir el estimador de prueba

$$R^{ts}(d) = \frac{1}{n_2}\sum(y_i - d(x_i))^2 \tag{2.11}$$

La última forma es construyendo el estimado de validacón cruzada de v subconjuntos (*v-fold*), que se deriva de dividir $\mathcal{L}$ en v subconjuntos del mismo tamao. Para cada subconjunto $\mathcal{L}_v$ se utiliza como muestra de aprendizaje a $\mathcal{L} - \mathcal{L}_v$ obteniendo el predictor $d^v(x)$, y a $\mathcal{L}_v$ para construir el estimador. Se obtiene entonces

$$R^{cv}(d) = \frac{1}{n}\sum_v \sum_{\mathcal{L}_v}(y_i - d^v(x_i))^2 \tag{2.12}$$

Nótese que el valor de $R^*(d)$ depende de la escala de medición de las variables respuesta, lo que provoca que su interpretación no sea del todo intuitiva. Por esto se puede construir una medida de eficacia en la que se elimine esta dependencia de la escala. Para hacerlo se utiliza $\mu = E(Y)$. Ahora, $R^*(\mu) = E(Y - \mu)^2$ es el error cuadrático medio cuando μ es usado como predictor de Y, y también es la varianza de Y. Utilizando esta estadística se puede construir el error cuadrático medio relativo en $d(x)$ como

$$RE(d) = \frac{R^*(d)}{R^*(\mu)} \tag{2.13}$$

Nótese que μ es el predictor de Y cuando no tenemos información sobre las variables explicativas X. $RE(d)$ compara el desempeo del predictor de X comparando su error cuadrático medio con el de μ. El valor $1 - RE(d)$ es llamado la proporción de varianza explicada por d cuando $d(x)$ es el mejor predictor lineal. De hecho cuando ρ es la coorrelación muestral entre los valores y_i y sus valores predichos $d(x_i)$ para $i = 1, ..., n$, entonces $\rho^2 = 1 - RE(d)$. En regresión lineal, cuando el predictor es el que ya conocemos ($d(x) = \beta_0 + \sum \beta_i x_i$). Éste valor es el que se conoce como R^2.

2.3.2. Construcción del árbol de regresión

Se revisará brevemente la construcción de los árboles de regresión, pues básicamente se sigue el mismo proceso que para los árboles de clasificación. La idea de trabajar con la estructura de un árbol para construir al predictor de la variable respuesta Y es similar a la utilizada cuando se construyen los árboles de clasificación: particionar recursivamente

el espacio muestral hasta alcanzar nodos terminales t, en los que se determina un valor predicho $y(t)$ de la variable respuesta. En cada nodo terminal t, el valor predicho se determina como constante.

De nuevo, se comienza con una muestra de aprendizaje, y los pasos para determinar al predictor son los mismos que para clasificación, a saber

- un método para elegir las divisiones en cada nodo intermedio
- una regla para determinar cuando un nodo es terminal
- una regla para asignar un valor y(t) a cada nodo terminal

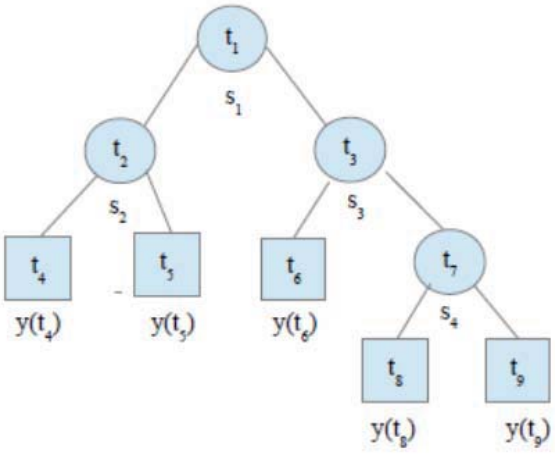


FIGURA 2.3: Ejemplo 3

Esta última parte es la más sencilla, pues simplemente se toma al estimado de resustitución para $R^*(d)$, y se elige el valor de y(t) que minimiza a $R(d)$, que es $y\hat{(}t)$ (la media de las y_i en el nodo t). Para cada nodo terminal t, el valor que le será asignado es la media $y\hat{(}t)$. Ahora, denotando a $R(d)$ como $R(T)$, sean

$$R(t) = \frac{1}{n}\sum_{i\in t}(y_i - y\hat{(}t))^2$$

$$R(T) = \sum_{t\in\tilde{T}} R(t)$$

Para cada nodo t, $nR(t)$ es la suma de las desviaciones al cuadrado de los y_i's en t de su media. Sumando sobre todos los $t \in \tilde{T}$, donde $\tilde{T}$ es el conjunto de los nodos terminales, se obtiene la suma de cuadrados total.

Para encontrar la mejor división s de un nodo t la idea es la misma a la utilizada en clasificación. Lo que se busca es una división s en S de tal forma que el decrecimiento en $R(T)$ sea máximo. Para formularlo con más precisión se puede tomarΔ $R(s,t) = R(t) - (R(t_L) + R(t_R))$, donde s divide a t en t_L y t_R. La mejor división s* será aquella que cumpla que

$$\Delta R(s*,t) = max_{s \in S} \Delta R(s,t)$$

Para determinar cuando un nodo es terminal, al igual que en clasificación se declara un número $\beta > 0$ arbitrario. Un nodo t será terminal cuando $max_{s \in S} \Delta R(s,t) < \beta$. $R(T)$ y $1 - RE(T)$ pueden ser utilizados como medidas de eficacia, pero el problema sigue siendo el mismo que cuando clasificamos: mientras más divisiones tengamos, el valor de $R(T)$ será menor, pero ésto no implica que el árbol sirva mejor como predictor. Lo que se hace entonces es continuar dividiendo hasta tener nodos terminales puros, donde todos los valores de y sean iguales (condición que rara vez se satisface) o hasta que el número de observaciones en el nodo sea menor que un número previamente determinado. Se obtiene así un árbol máximo T_{max}, que iremos podando, es decir eliminando divisiones generando árboles más pequeos, hasta encontrar uno de tamao óptimo. El proceso para hacerlo se basa en la medida de error-complejidad que ya fue mencionada anteriormente.

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

Se obtiene así una secuencia de árboles $T_max > T_1 > ... > t_1$ donde t_1 es el árbol mínimo (aquel que solo contiene al nodo raíz), y una secuencia correspondiente $0 = \alpha_1 < \alpha_2 < ...$ tal que para $\alpha_k < \alpha < \alpha_{k+1}$, T_k es el mínimo subárbol de T_{max} que minimiza a $R_\alpha(T)$. Análogamente a como se selecciona al árbol óptimo de clasificación, se selecciona al árbol óptimo de regresión. Para todos los valores de α se encontrará al árbol óptimo mediante la medida resustitución de costo-complejidad, $R_\alpha(T)$. De éstos, el mejor árbol podado es el correspondiente al valor α_0 tal que

$$R_{\alpha 0}(T) \leq R_\alpha(T)$$

para todos los valores de α.

Capítulo 3

Árboles de regresión logística: algoritmo LOTUS

A pesar del uso que se le puede dar a la regresión logística, ésta no está exenta de problemas. Por ejemplo, cuando se tienen presentes muchas variables explicativas puede existir colinearidad o interacciones entre éstas que no permitan una interpretación sencilla o intuitiva del modelo, algo que sí es posible hacer cuando se ajustan regresiones para una sola variable explicativa, como se vió en el capítulo 1. Otro problema es que no existen buenos métodos para juzgar la bondad de ajuste del modelo, y que permitan saber cuales son las transformaciones adecuadas que se deben aplicar a las variables. Ahora que se ha desarrollado tanto la teoría de regresión logística como la de árboles de clasificación, se utilizará el algoritmo LOTUS (*Logistic Tree with Unbiased Selection*) desarrollado por Chan y Loh (2004) en el que se juntan estas dos técnicas para generar modelos en los que se resuelven los problemas anteriores. Dicho algoritmo genera un árbol a partir de divisiones apropiadas. Una vez obtenidos los nodos terminales, se aplica un modelo de regresión logística a cada uno de ellos. La ventaja es que al utilizar solamente regresiones con una variable independiente, la interpretación -que parte del árbol y culmina con la regresión- es muy intuitiva. Además, el algoritmo LOTUS permite poder modelar relaciones no lineales sin requerir transformar a las variables, algo que complica mucho la interpretación.

Se sabe que la regresión logística modela la relación entre la probabilidad $P(Y = 1)$ y un conjunto de variables X independientes a través de la función logit,

$$logit(x) = ln\frac{\pi(x)}{1 - \pi(x)} \tag{3.1}$$

donde $\pi(x) = \beta_0 + \sum_i \beta_i x_i$.

La forma de proceder del algoritmo LOTUS parte de particionar el espacio muestral mediante la generación de un árbol. Los árboles de clasificación tratan de predecir la pertenencia a una clase j donde $j \in \{1, ..., J\}$. En este caso, es posible pensar en el problema como un caso particular donde sólo se tienen dos clases, una para $Y = 1$ y otra para $Y = 0$. Entonces, desde este punto de vista, sería viable pensar que una vez alcanzado un nodo terminal, la asignación de clase se puede hacer como se hacía con el algoritmo CART: utilizando proporciones muestrales. Si se denota como $n_1 = \sum y_i$ a la suma de observaciónes tales que $y_i = 1$, y como $n_0 = \sum y_i$ a la suma tal que $y_i = 0$, y como n al número total de observaciones, el algoritmo CART elige el máximo entre n_1 y n_0 y asigna al nodo la clase correspondiente a dicho máximo. Esta selección de clases tiene dos problemas, la clase estimada es constante en cada nodo, y la estimación es precisa solamente para árboles muy grandes, que son muy difíciles de interpretar debido al número de particiones. Al ser el objetivo facilitar la interpretación, los árboles muy grandes no son adecuados. El modelo LOTUS soluciona el problema con la aplicación de un modelo logístico en cada uno de los nodos terminales. Además de que permite una interpretación clara, el modelo del algoritmo LOTUS tiene también la ventaja de que el sesgo al elegir las divisiones de los nodos ya no se presenta. Además, es aplicable tanto a variables categóricas como cuantitativas.

3.1. Predictores

Las variables independientes que se utilizan para predecir la variable de respuesta pueden tomar tres papeles en el modelo LOTUS. Si son utilizadas solamente para generar las divisiones serán llamadas variables s. Si por el contrario están restringidas a actuar como regresoras en el ajuste logístico, se les llama variables f. Si se permite que actúen tanto como regresoras como en la partición, se les llama variables n.

Para una variable X_i, l_m y l_s denotan la log-verosimilitud maximizada para los modelos de interés (el ajustado y el saturado, respectivamente). Como se ha mencionado anteriormente, la devianza es igual a $D = -2(l_m - l_s)$, y es una estadística análoga a la suma de cuadrados de residuales. Para predecir la clase (Y=1 o Y=0) a la que pertenecen las observaciones de un nodo, se debe elegir entre las variables f y n a aquella que, al ajustar una regresión logística, arroje el modelo con menor devianza por grados de libertad (# de observaciones ajustadas $-$# de parámetros estimados, incluyendo el parámetro de intercepción). Este valor mide la impureza del nodo. La impureza total de un árbol está dada por la suma de las devianzas en cada nodo terminal. En resumen, si t es un nodo

terminal y $X_1, ..., X_K$ son las variables categorizadas como variables f ó n en dicho nodo, procedemos de la siguiente manera:

- ajustar el modelo
$$\beta_0 + \beta_1 x_k = ln\frac{\pi(x)}{1-\pi(x)}$$
para todo $k = 1, ..., K$.

Si D_k es la devianza para el modelo k y v_k los grados de libertad correspondientes, denotamos como

$$\hat{D_k} = \frac{D_k}{v_k}$$

al cociente entre la devianza y los grados de libertad del modelo correspondiente a X_k, para cada variable X del tipo f ó n. Si denotamos como k^* a aquel valor de k que minimice a $\hat{D_k}$, entones se elegirá para ese nodo t el modelo $\beta_0 + \beta_1 x_k^*$. En el caso de que $\hat{D_k} = \infty$, lo que puede suceder cuando los datos son puros con respecto a la variable Y o bien si el modelo no converge, se elimina la división correspondiente en dicho nodo.

3.2. Selección de variables para la división

En esta sección se revisa el problema de cómo elegir las variables para dividir los nodos. Generalmente en los algoritmos en los que se lleva a cabo una búsqueda exhaustiva para seleccionar una variable para generar una partición, suele haber un sesgo para elegir variables cuyo número posible de particiones es grande. Para solucionar ésto, el algoritmo LOTUS parte el proceso de dividir los nodos: primero elige entre las variables s y n a la más adecuada (siguiendo los criterios que se detallan a continuación), y después se encuentra el valor o el subconjunto de X para particionar al nodo. La justificación para seguir este proceso se sigue de que el algoritmo CART tiene generalmente problemas de sesgo, es decir, se da preferencia a las variables independientes que tienen un número muy grande de posibles divisiones.

Se ha observado que el algoritmo CART suele elegir primero entre variables en las que el número de divisiones posibles es muy grande (Loh y Chan, 2002). Si dichas variables realmente fueran las más importantes para explicar el comportamiento de Y, se esperaría que el costo de error de clasificación (al que se le llama $R^{cv}(d)$) se incremente de forma significativa si se elimina a estas variables del modelo, lo cual no sucede. Una forma de evitar el problema es incluir un factor de ponderación, pero ésto suele ser

computacionalmente poco eficiente y no hay una forma clara de hacerlo. Debido a ésto, el algoritmo LOTUS primero elige a la mejor variable y después el valor c de división.

El algoritmo que utilizan Chan y Loh retoma la idea del modelo GUIDE (*Generalized, Unbiased, Interaction Detection and Estimation*, Loh, 2002), ajustándolo para regresiones logísticas. A continuación detallaremos como procede el modelo GUIDE para después ver como se ajusta para ser aplicado en LOTUS.
La idea del algoritmo GUIDE es usar pruebas de curvatura. Dichas pruebas siguen en cada nodo los siguientes pasos:

1. Se ajusta a las observaciones correspondientes al nodo un modelo constante, a saber la media muestral $\bar{Y}$.
2. Se detecta el signo del residual $Y - \hat{Y}$, con lo que se generan dos grupos, uno con los residuales positivos y el otro con los negativos.
3. La variable con menor valor-p entre las pruebas t para dos muestras y la prueba de Levine para medias y varianzas desiguales respectivamente se selecciona para dividir al nodo.

El método anterior tiene dos problemas, (1) no es aplicable a variables independientes categóricas y (2) no detecta interacción entre variables independientes. El primer problema se soluciona utilizando la prueba χ^2 de Pearson, que detecta la asociación entre los residuales y los grupos de predictores. A continuación se detalla dicha prueba.

3.2.1. Prueba χ^2 de Pearson

La prueba de χ^2 de Pearson consiste en calcular que tan probable es que las diferencias observadas en una tabla de contingencia entre subconjuntos de la variable X sean aleatorias. La estdaística correspondiente es:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3.2}$$

donde O_i es la frecuencia observada en la celda i, E_i la frecuencia esperada, y n el número de celdas. Los grados de libertad correspondientes son $n - p$, con p la reducción en grados de libertad. El valor calculado de la estadística se compara entonces con los cuantiles de una distribución χ^2_{n-p}.

Para variables categóricas, la prueba se aplica sobre una tabla de contingencia de $2 \times c$, con el signo del residual $Y - \hat{Y}$ en las filas y las categorías de la variable en las columnas. En el caso de variables numéricas, el algoritmo del modelo GUIDE las vuelve discretas generando cuatro grupos a partir de los cuartiles muestrales, obteniendo una tabla de 2×4. En ambos casos se obtiene la estadística χ^2 y los valores-p correspondientes, ya sea para una χ^2 con $c-1$ o $4-1$ grados de libertad, correspondientes al tamaño de las tablas.
Para detectar interacciones entre variables, se considera el espacio generado por cada pareja de variables (X_i, X_j). En el caso de que tanto X_i como X_j sean variables numéricas, se divide el espacio en cuatro cuadrantes, tomando a las medias muestrales para particionarlo. Se genera entonces una tabla de contingencia de 2×4. En el caso de dos variables categóricas, la tabla será de $2 \times c_i c_j$, y para una categórica y una numérica, de $2 \times 2c$. En los tres casos se aplica de la misma forma la prueba de la Ji-Cuadrada de Pearson. Dichas pruebas se llamarán pruebas de interacción.
Por último, se procede a comparar entre todos los valores-p anteriores. Si el menor corresponde a una prueba de curvatura, se selecciona a la variable correspondiente para particionar al nodo. Si por otro lado el menor valor-p corresponde a una prueba de interacción, se selecciona de las dos variables correspondientes a aquella que reduzca más a SSE, la suma de cuadrados residuales.

El algoritmo anterior solo es aplicable a modelos lineales. La idea detrás de las pruebas de curvatura, es decir pruebas ji-cuadradas de asociación sobre tablas de contingencia, es que si el modelo se ajusta bien, sus residuales deben tener mínima asociación con los valores de la variable predictora y el valor-p de la prueba χ^2 se debe distribuir uniformemente. Por otro lado, si el efecto de la variable no es tomado en cuenta, el valor-p se espera que sea pequeño. Entonces, la forma de elegir a la variable para la división es elegir a la de menor valor-p. Aunque el proceso no es aplicable directamente a la regresión logística, puesto que el método no distingue entre efectos lineales y no lineales de X sobre Y. Para distinguir los efectos lineales de los no lineales, se usa una prueba Ji-cuadrada para tendencias (Cochran 1954 y Armitage 1955).
El problema reside en que, en regresión logística, el residual $Y - \hat{p}$ es siempre positivo. En este caso, llevar a cabo una prueba de asociación entre los residuales y los valores de una variable independiente X corresponde simplemente a una prueba de asociación entre Y y X. Al ser la prueba independiente del modelo elegido, ésta no detecta si el efecto de la interacción de X sobre Y es lineal, que es precisamente lo que se quiere. Para solucionar el problema se genera una tabla de $2 \times J$, donde las columnas son los valores de Y (a saber 0 o 1) y las filas los valores de $X = x_j$, con $j = 1, ..., J$. Se le llama n_{ij} al número de casos con $Y = i$ y $X = x_j$. El tamaño de muestra entonces es $n = \sum_i \sum_j n_{ij}$ y la proporción muestral en cada celda $p_{ij} = n_{ij}/n$. La proporción condicional de observaciones en la

columna j con respuesta i será $p_{i|j} = p_{ij}/p_{+j} = n_{ij}/n_{+j}$ donde $n_{+j} = np_{+j} = \sum_i n_{ij}$. Para un modelo logístico lineal $\pi_{1|j} = \beta_0 + \beta_1 x_j$ donde $\pi_{i|j}$ es la probabilidad de que Y=1 dado el valor de $X = x_j$ se ajusta por mínimos cuadrados obteniendo $\hat{\pi_{1|j}}$. La estadística χ^2 de Pearson para independencia se descompone como:

$$\chi^2 = z^2 + \chi_L^2 \tag{3.3}$$

donde z^2 es la estadística de prueba para probar tendencia de Cochran-Armitage, y es una prueba para tendencia lineal en las proporciones. Lo que se busca es ver si el efecto de X sobre Y es lineal, es decir verificar si el modelo $\pi_{1|j} = \beta_0 + \beta_1 x_j$ es válido. Cuando el modelo se sostiene, z^2 tiene una distribución asintótica Ji-cuadrada con un grado de libertad, $\chi^2_{(1)}$. Por otro lado la estadística χ_L^2 tiene una distribución asintótica Ji-cuadrada con J-2 grados de libertad, $\chi^2_{(J-2)}$ y prueba independencia entre X y Y después del ajuste para tendencia lineal. Para variables numéricas se obtienen los valores-p, después de discretizarlos en los cinco quintiles. En el caso de las variables categóricas, al solo ser utilizadas para dividir los nodos, basta usar la prueba χ^2 para verificar interacción entre Y y X. El valor-p se obtiene con una χ^2 con $c(t) - 1$ grados de libertad, donde $c(t)$ corresponde al número de categorías de la variable X en el nodo t. Después de obtener el valor-p para cada variable s, n, c y o (las candidatas para dividir al nodo) la de menor valor-p es utilizada para dividir el nodo.

3.3. Selección del punto de división

El algoritmo CART elige el punto de división s^* de tal forma queΔ $R(s^*, t) = R(t) - (R(t_L) + R(t_R)) = max_{s \in S} \Delta R(s, t)$, donde s divide a t en t_L y t_R.

Si se piensa en implementar un método análogo cuando la variable elegida es cuantitativa, el punto de división c será aquel que minimice la devianza total del modelo de regresión logística ajustado comparado con la de los dos nodos generados por la división. El algoritmo LOTUS utiliza un método más sencillo, restringiéndose a los cuantiles muestrales de X.

Si la variable X es categórica, también se limita a elegir cinco candidatos. Si $a_1, ..., a_m$ es el conjunto de valores posibles de X en el nodo t y $q(a)$ es la proporción de casos con $Y = 1$ entre aquellos con $X = a$, y si $a_{(1)}, ..., a_{(m)}$ es el conjunto de valores ordenados de a de acuerdo a $q(a)$ (es decir $q(a_{(1)}) \leq ... \leq q(a_{(m)})$) se dice que A^* minimiza la suma de la varianza en Y dado X. La división a partir de los $X \in A^*$ se puede encontrar buscando sobre las $m - 1$ divisiones X_i donde $A_i = a_{(1)}, ..., a_{(i)}$ con $i = 1, ..., m - 1$. Si se denota a $A^* = a_{(1)}, ..., a_{(j)}$ con $j = \pm 1, \pm 2$ de tal forma que

$1 + j - 1$ y se define a $A*_j = a_{(1)}, ..., a_{(l+j)}$, LOTUS evalúa las cinco divisiones $X \in A$ con $A = A^*_{-2}, A^*_{-1}, A^*, A^*_1, A^*_2$ y elige aquella que minimice la suma de las devianzas logísticas en los dos subnodos o nodos hijos, al igual que para variables cuantitativas. Si ningún valor de A es adecuado se busca a la siguiente variable X entre las elegidas para dividir y se hace su división correspondiente. Ésto se hace para evitar una terminación prematura del crecimiento del árbol.

3.4. Poda del árbol

Para la poda del árbol se implementa el método de mínimo costo de complejidad del algoritmo CART (más detalles se muestran en el capítulo anterior), donde el costo es la devianza predicha estimada. Si hay una muestra de prueba independiente se puede utilizar para calcular el estimador de resustitución. De otra forma, se hace una validación cruzada con diez subconjuntos, que fué denotado como $R^{cv}(d)$ en el capítulo anterior. El subárbol con menor devianza predicha estimada será el elegido.

Si se revisa a detalle el algorítmo LOTUS, es posible darse cuenta que el valor del trabajo de Chan y Loh no consiste tanto en la novedad, sino en el haber logrado conjuntar de forma clara y funcional métodos que por si solos son eficientes, pero que al mezclarse logran evitar varios de sus problemas, y así generar un modelo de fácil construcción e interpretación. Como se menciona en el capítulo, el algorítmo LOTUS no es el primero que junta la regresión logística y la estructura de los árboles. Ésto es un claro indicador de que tanto uno como el otro son modelos que no están excentos de problemas, siendo los más directos la difícil interpretación de modelos con muchas variables explicativas, y la selección sesgada de variables, que llevan a conclusiones incorrectas. LOTUS es solo un ejemplo de que cuando se trabaja juntando diferentes herramientas es posible obtener modelos precisos sin sacrificar su interpretación.

Capítulo 4

Ejemplos: Algoritmo CART

4.1. Ejemplo 1: Pacientes con síndrome coronario (*Evans County Dataset*)

Se retoma el ejemplo de los datos del Condado de Evans (Evans County Dataset), que presenta los datos de un estudio en el que se siguió a 609 pacientes masculinos durante 7 años, siendo la presencia de síndrome coronario (que se denota CHD) la variable de interés. La descripción de las variables se encuentra en el capítulo 1. Para analizar los datos e implementar el algoritmo CART, se utiliza el software estadístico 'R' y el paquete 'rpart'. La función 'summary' presenta información sobre la muestra: los cuantíles más importantes, la media y mediana, así como el valor máximo y mínimo para cada variable. En el caso de variables categóricas, se muestra el número de observaciones en cada categoría. En este ejemplo, todas las variables categóricas son binarias.

Cuadro 4.1: Resumen de Variables: Evans County Dataset

ID	CHD	CAT	AGE	CHL
Min. : 21	NO :538	0:487	Min. :40.00	Min. : 94.0
1st Qu.: 4242	YES: 71	1:122	1st Qu.:46.00	1st Qu.:184.0
Median : 9751			Median :52.00	Median :209.0
Mean : 9213			Mean :53.71	Mean :211.7
3rd Qu.:13941			3rd Qu.:60.00	3rd Qu.:234.0
Max. :19161			Max. :76.00	Max. :357.0
SMK	ECG	DPB	SPB	
0:222	0:443	Min. : 60.00	Min. : 92.0	
1:387	1:166	1st Qu.: 80.00	1st Qu.:125.0	
		Median : 90.00	Median :140.0	
		Mean : 91.18	Mean :145.5	
		3rd Qu.:100.00	3rd Qu.:160.0	
		Max. :170.00	Max. :300.0	

Se puede, por ejemplo, comparar a los pacientes con y sin la enfermedad con una gráfica de caja. Se observa que la edad promedio de los pacientes en el estudio es de 52 años, siendo los valores en el primero y tercer cuartíles 46 y 60 años respectivamente. Ahora, para pacientes con CHD, la media sube a 57.25 y para aquellos sin la enfermedad desciende a 53.24.

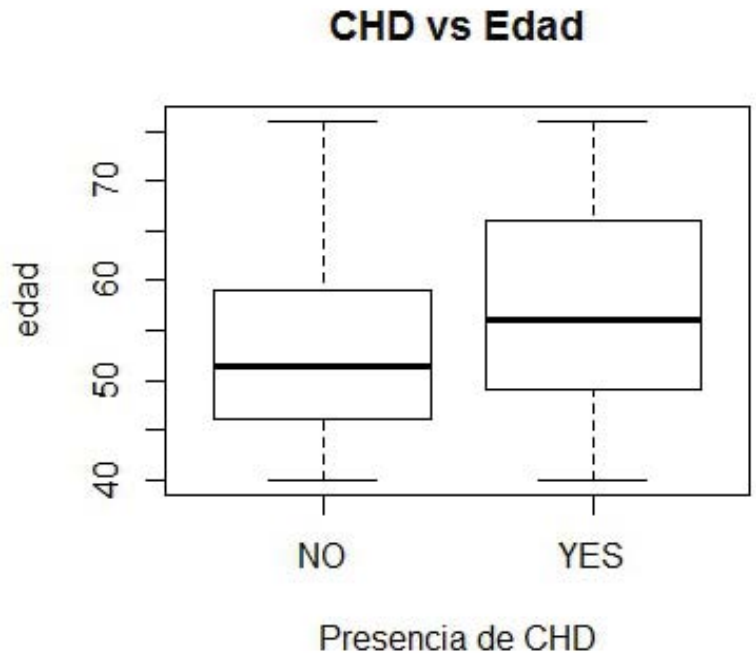


Figura 4.1: Gráfica de caja: CHD vs AGE

Para las variables categóricas, las tablas de frecuencias revelan información sobre la distribución de las observaciones. Por ejemplo, de los pacientes con CHD, 54 fumaban y 17 no.

Cuadro 4.2: Tabla 4.2. Tabla de frecuencias para la variable SMK

	SMK=0	SMK=1
CHD=YES	205	333
CHD=NO	17	54

Para las variables HBP y CAT, que representan presión alta y niveles de catecotamína, se hace lo mismo, obteniendo las siguientes tablas:

CUADRO 4.3: Tablas de frecuencias y porcentajes para la variable HBP

Frecuencia	HBP=0	HBP=1
CHD=YES	326	212
CHD=NO	28	43

Porcentaje	HBP=0	HBP=1
CHD=YES	0.921	0.83
CHD=NO	0.079	0.17

Es posible observar que de los pacientes con niveles normales de presión (HBP=0), el 92 % no presentaron la enfermedad.

CUADRO 4.4: Tablas de frecuencias y porcentajes para la variable CAT

Frecuencia	CAT=0	CAT=1
CHD=YES	443	95
CHD=NO	44	27

Porcentaje	CAT=0	CAT=1
CHD=YES	0.90	0.78
CHD=NO	0.1	0.22

Por otro lado, entre los pacientes con altos niveles de catecotamína, el 22 % tienen presencia de CHD, mientras que 71 de 609, es decir solo el 11 %, presentan CHD.
Los histogramas muestran gráficamente la distribución de las observaciones. Para la edad, se observa que la mayoría de los pacientes esta entre 40 y 55 años de edad. Por otro lado la distribución de los niveles de colesterol parece seguir una distribución Normal alrededor de 211. Para pacientes con CHD sube a 221, lo que parece relevante si se nota que para el primero y tercer cuartíles los valores son 184 y 234.

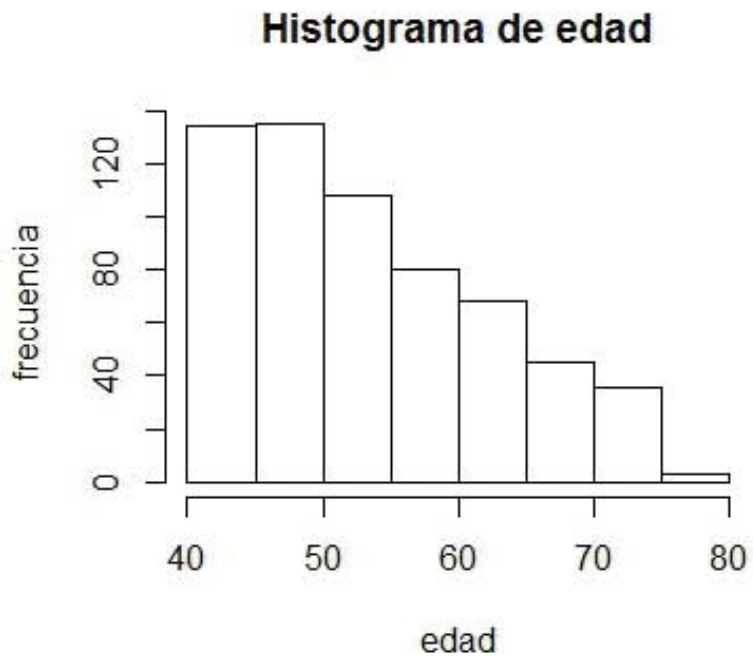


FIGURA 4.2: Histograma: Edad

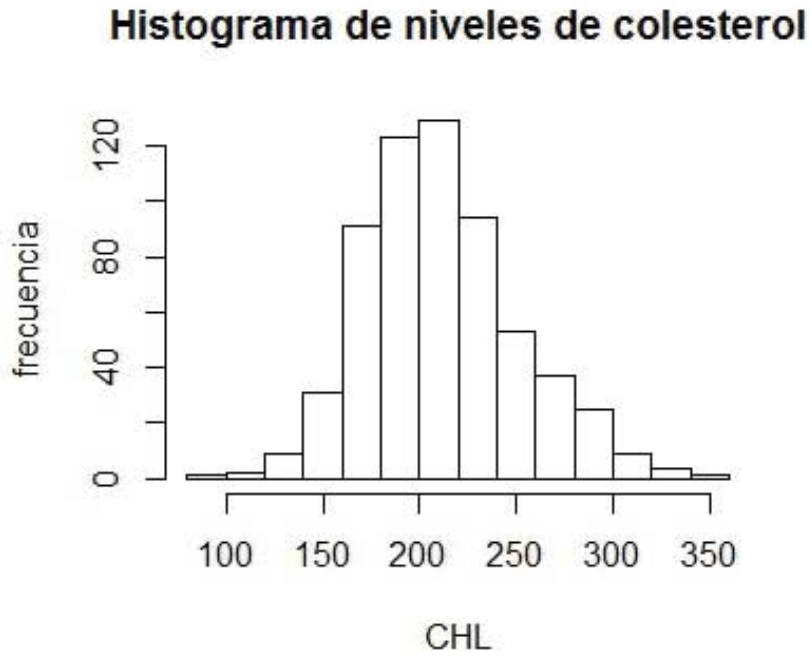


FIGURA 4.3: Histograma: Nivel de Colesterol

Se procede a modelar los datos mediante un árbol de clasificación. Las variables que son utilizadas son CHD como variable respuesta y CAT, AGE, CHL, SMK, ECG y HPT como variables explicativas. Nótese que hay cuatro variables categóricas, que de hecho son indicadoras, y dos variables cuantitativas: la edad y el nivel de colesterol. El ajuste se hace mediante la función 'rpart', y es mostrado con la función 'print'.

```
> ajuste<- rpart(CHD~ CAT+AGE+CHL+SMK+ECG+HPT,data=evans,method="class")
> print(ajuste)
n= 609
node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 609 71 NO (0.88341544 0.11658456)
   2) AGE< 62.5 474 40 NO (0.91561181 0.08438819) *
   3) AGE>=62.5 135 31 NO (0.77037037 0.22962963)
     6) CHL< 229.5 96 15 NO (0.84375000 0.15625000)
      12) CHL>=213.5 24  1 NO (0.95833333 0.04166667) *
      13) CHL< 213.5 72 14 NO (0.80555556 0.19444444)
        26) CAT=1 39  3 NO (0.92307692 0.07692308) *
      27) CAT=0 33 11 NO (0.66666667 0.33333333)
         54) HPT=0 21  1 NO (0.95238095 0.04761905) *
          55) HPT=1 12  2 YES (0.16666667 0.83333333) *
     7) CHL>=229.5 39 16 NO (0.58974359 0.41025641)
    14) CAT=0 22  3 NO (0.86363636 0.13636364) *
      15) CAT=1 17  4 YES (0.23529412 0.76470588) *
```

La información mostrada es la siguiente:

- tipo de nodo o forma de la división.
- número de observaciones.
- error en el nodo.
- categoría.

En el caso del error del nodo, se usan probabilidades a priori. Por ejemplo, en el nodo raíz hay 71 observaciones con CHD='NO'. Esos 71 son los mal clasificados en este primer nodo. También se puede observar que los nodos marcados con un asterisco (*) son los nodos terminales. Para cada nodo llamado n, sus subnodos o nodos hijos son los nodos $2n$ y $2n+1$, donde el de la izquierda siempre es $2n$. Así, por ejemplo, el nodo raíz es 1), y sus subnodos 2) y 3). En ese caso las observaciones en el subnodo izquierdo son aquellas en las que $AGE < 62.5$, de las cuales tenemos 474. De éstas, 40 están mal clasificadas,

es decir presentan CHD, pues también se nos indica que la categoría asignada es 'NO'. Por otro lado, el subnodo derecho es aquel en el que $AGE \geq 62.5$. La lectura de la información se hace de misma forma. La figura 4.4 presenta gráficamente los resultados, lo que puede clarificar la información mostrada y permitir interpretar el árbol.

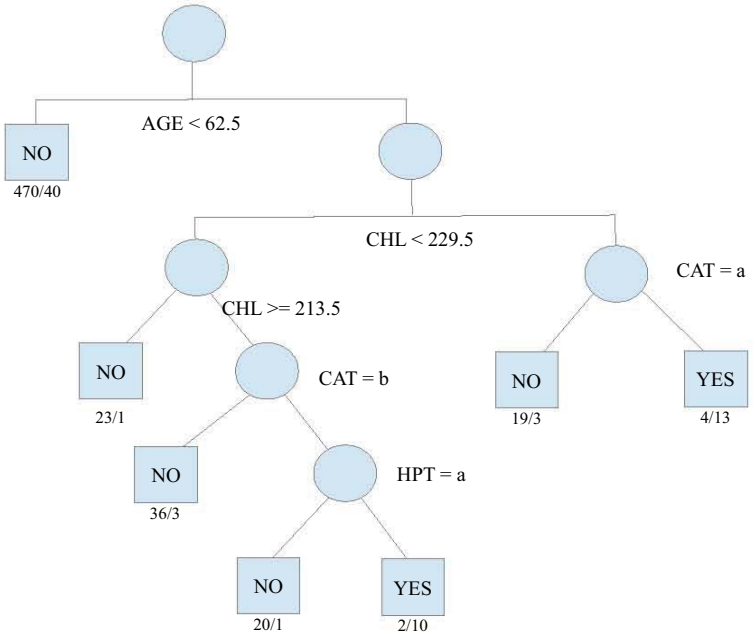


FIGURA 4.4: Árbol de clasificación: Evans County Dataset

Para analizar los resultados, el paquete 'rpart' provee también a la función 'printcp',

que muestra una tabla con podas óptimas para los correspondientes parámetros de complejidad. Para todo subárbol T, se definió su complejidad como $|\tilde{T}|$, que es el número de nodos terminales. Se define también a un número real $\alpha \geq 0$ como el parámetro de complejidad, y la medida costo-complejidad como:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

que es una combinación lineal del costo del árbol y su complejidad. Si para cada valor de α se encuentra el subárbol que minimice a $R_\alpha(T)$, es decir,

$$R_\alpha(T(\alpha)) = min_{T<T_{max}} R_\alpha(T)$$

Corriendo sobre los valores de α, se genera una sucesión de subárboles $T_1, T_2, ..., T_{max}$, con cada vez menos nodos terminales. El problema de elegir el mejor subárbol es un problema de optimización y es discutido en el capítulo 3.4 de Breiman [1]. Ahora, la tabla generada en 'R' presenta la siguiente información:

- Root node error: proporción de errores de clasificación en el nodo raíz, de la forma n_1/n donde n es el total de observaciones.
- CP: parámetro de complejidad.
- nsplit: número de particiones.
- relerror: error de clasificación, usando el estimador de prueba.
- xerror: error de clasificación, usando una muestra de prueba.
- xstd: error de clasificación, usando el método de validación cruzada.

Las últimas tres columnas muestran el valor cambio de escala para que el árbol T_{max} tenga el valor de 1. Para obtener el valor absoluto de errores de clasificación, se multiplica el valor dado por n_1. La tabla, en este ejemplo, arroja los siguientes resultados:

```
> printcp(ajuste)

Classification tree:
```

```
rpart(formula = CHD ~ CAT + AGE + CHL + SMK + ECG + HPT, data = evans,
    method = "class")

Variables actually used in tree construction:
[1] AGE CAT CHL HPT

Root node error: 71/609 = 0.11658

n= 609
```

	CP	nsplit	rel error	xerror	xstd
1	0.042254	0	1.00000	1.00000	0.11155
2	0.037559	3	0.87324	1.00000	0.11155
3	0.010000	6	0.76056	0.85915	0.10435

Por un lado, se observa que las variables SMK y ECG no fueron utilizadas para ajustar el modelo. Para los tres valores del parámetro de complejidad, se obtienen los correspondientes subárboles. En este caso, el subárbol con solo el nodo raíz, un subárbol con 3 divisiones, y el árbol máximo. El error de clasificación no decrece para los primeros dos, por lo que no es necesario hacer una poda: nos quedamos con el árbol máximo como el que mejor clasifica a las observaciones. Esto parece razonable debido a que en este ejemplo, el árbol máximo no tiene una cantidad muy grande de divisiones.
La función 'summary' presenta más información sobre los nodos. Primero se muestra un resumen sobre el peso o la importancia general de las variables. Como se observa, se encontró que ECG y SMK no explican muy bien la presencia de CHD. Es por esto que no fueron utilizadas para ajustar el árbol. Se cuenta también con la información extendida de cada nodo. Se puede observar por ejemplo, las divisiones que fueron candidato para generar los subnodos, así como el decrecimiento en impureza generado por cada una de éstas. También se muestra el conteo de observaciones correspondientes a cada categoría, el número de observaciones correspondientes a cada subnodo (cuando éstos están presentes) y la clase asignada. La información desplegada para los primeros nodos es la siguiente:

```
> summary(ajuste)

Variable importance
HPT CAT CHL AGE ECG SMK
 28  23  21  17   8   3

Node number 1: 609 observations,    complexity param=0.04225352
  predicted class=NO   expected loss=0.1165846  P(node) =1
    class counts:   538    71
   probabilities: 0.883 0.117
  left son=2 (474 obs) right son=3 (135 obs)
```

```
  Primary splits:
      AGE < 62.5  to the left,  improve=4.433084, (0 missing)
      CAT splits as  LR,        improve=3.346530, (0 missing)
      HPT splits as  LR,        improve=2.376331, (0 missing)
      ECG splits as  LR,        improve=1.541405, (0 missing)
      CHL < 199.5 to the left,  improve=1.523956, (0 missing)

Node number 2: 474 observations
  predicted class=NO   expected loss=0.08438819  P(node) =0.7783251
    class counts:   434    40
   probabilities: 0.916 0.084

Node number 3: 135 observations,    complexity param=0.04225352
  predicted class=NO   expected loss=0.2296296  P(node) =0.2216749
    class counts:   104    31
   probabilities: 0.770 0.230
  left son=6 (96 obs) right son=7 (39 obs)
  Primary splits:
      CHL < 229.5 to the left,  improve=3.5786680, (0 missing)
      AGE < 67.5  to the right, improve=1.1875660, (0 missing)
      HPT splits as  LR,        improve=1.0804960, (0 missing)
      CAT splits as  LR,        improve=0.6229630, (0 missing)
      SMK splits as  LR,        improve=0.5080179, (0 missing)
  Surrogate splits:
      AGE < 75.5  to the left,  agree=0.719, adj=0.026, (0 split)
```

4.1.1. Interpretación del árbol

El objetivo de modelar los datos es encontrar cómo afectan ciertas variables sobre la presencia de síndrome coronario. Siguiendo los resultados del ajuste, y observando la figura 4.4, es posible concluir que las variables que tienen más peso sobre la presencia o ausencia de CHD son la edad, el nivel de colesterol y el nivel de catecotamína. A diferencia de lo que se puede pensar antes de ver el modelo, se encuentra que la variable SMK no tiene gran importancia para explicar la presencia de CHD. Ahora, por otro lado, se encuentra que la edad sí es un factor importante para explicarla. Aquellos pacientes menores a 62.5 años son clasificados como pacientes de bajo riesgo. Ahora, para aquellos mayores a 62.5 años, la siguiente variable importante es el nivel de colesterol, que divide a estos pacientes asignando a aquellos con nivel menor a 229.5 al subnodo izquierdo y a aquellos con nivel mayor al derecho. De los pacientes con colesterol bajo, de un total de 96, 84 son clasificados como de bajo riesgo y solo aquellos con nivel alto de catecotamína y presión cardiaca alta (12) se consideran como de alto riesgo. De estos 96, hay sólo 7 mal clasificados, lo que corresponde aproximadamente a un 7.2 %. Ahora, para los pacientes con colesterol alto (mayor a 229.5), incluso aquellos con nivel de catecotamína

normal son clasificados como de bajo riesgo de contraer CHD. Recuérdese que la media del nivel de colesterol es de 211.7. Entonces, se puede deducir que incluso para los pacientes mayores de edad, los que tienen un nivel moderado de colesterol y niveles normales de catecotamína tampoco son pacientes con alto riesgo de contraer síndrome coronario. Ahora, si se considera a los 474 pacientes menores de 62.5 años, hay 40 mal clasificados. Ésta es una de las posibles fallas del modelo CART, pues a pesar de que la frecuencia observada de pacientes con CHD es del 11 %, y en este nodo se clasifica mal solamente al 8.4 %, aún así, en números absolutos, clasificar mal a 40 de 474 pacientes con CHD es un valor alto. Una posible solución a este problema puede ser ajustar un nuevo modelo para los pacientes menores a 62.5 años, ya sea un modelo logístico, u otro árbol de clasificación. Excluyendo a estos pacientes, el árbol de clasificación hace un buen trabajo en explicar la influencia de las variables sobre la enfermedad. Las variables relevantes son el nivel de colesterol, el de catecotamina, y la presión arterial en menor medida. Se concluye que incluso para pacientes mayores, aquellos con niveles adecuados en las variables anteriores no corren alto riesgo de contraer síndrome coronario.

4.2. Ejemplo 2: Cangrejos

Retomando un ejemplo presentado previamente, se modelarán los datos de los cangrejos herradura *(horseshoe crabs)*, ajustaremos ahora un árbol de clasificación mediante el algoritmo descrito previamente, que sigue el modelo CART. Recordemos que el problema es clasificar a los cangrejos dentro de dos clases: aquellos con satélites y aquellos sin éstos. Utilizaremos esta vez las variables de color, peso, condición de la espina y ancho, como variables independientes. De éstas, tenemos dos categóricas (color y condición de la espina), mientras que las restantes son cualitativas. Un primer ajuste arroja los siguientes resultados:

```
n= 172

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 172 62 1 (0.3604651 0.6395349)
   2) width< 25.85 78 34 0 (0.5641026 0.4358974)
     4) color=4,5 37 11 0 (0.7027027 0.2972973) *
     5) color=2,3 41 18 1 (0.4390244 0.5609756)
      10) width>=25.55 8  2 0 (0.7500000 0.2500000) *
      11) width< 25.55 33 12 1 (0.3636364 0.6363636)
        22) width< 24.95 20 10 0 (0.5000000 0.5000000)
          44) weight>=1975 8  2 0 (0.7500000 0.2500000) *
```

```
        45) weight< 1975 12  4 1 (0.3333333 0.6666667) *
      23) width>=24.95 13  2 1 (0.1538462 0.8461538) *
  3) width>=25.85 94 18 1 (0.1914894 0.8085106)
    6) color=5 7  2 0 (0.7142857 0.2857143) *
    7) color=2,3,4 87 13 1 (0.1494253 0.8505747) *
```

Podemos obtener también una gráfica del árbol:

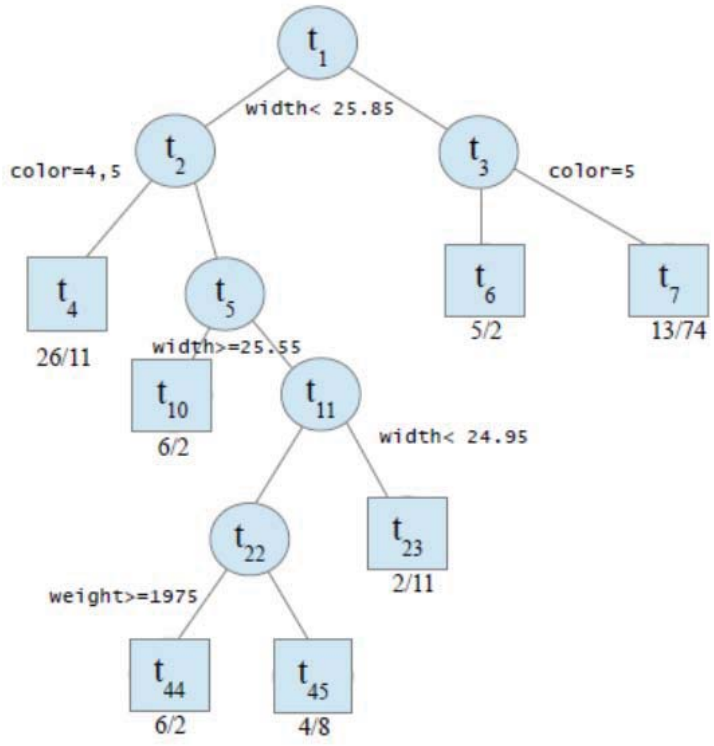


FIGURA 4.5: Árbol de clasificación: Horseshoe Crabs

La tabla, en nuestro ejemplo, arroja los siguientes resultados:

```
Classification tree:
rpart(formula = Y ~ color + spine + width + weight, data = datos,
    method = "class")

Variables actually used in tree construction:
[1] color  weight width

Root node error: 62/172 = 0.36047

n= 172
```

```
        CP nsplit rel error  xerror    xstd
1 0.161290      0   1.00000 1.00000 0.10156
2 0.080645      1   0.83871 1.09677 0.10342
3 0.064516      2   0.75806 1.03226 0.10225
4 0.048387      3   0.69355 1.03226 0.10225
5 0.032258      4   0.64516 1.01613 0.10191
6 0.010000      6   0.58065 0.98387 0.10120
```

El nodo raíz clasifica correctamente solamente 62 de 172 observaciones. Aparentemente, hacer un ajuste con estos datos no arroja buenos resultados. En el siguiente capítulo, se ajustará un modelo mediante el algoritmo LOTUS, de forma que será posible comparar entre ambos.

Capítulo 5

Ejemplos: algoritmo LOTUS

Para ejemplificar como funciona el algoritmo LOTUS, se tomarán dos ejemplos. Primero se utilizarán los datos sobre la presencia de satélites en cangrejos hembra, con los que ya se ha trabajando previamente. Después se revisarán los datos sobre coches producidos en EUA entre los años 1970 y 1982, tomados de StatLib (http://lib.stat.cmu.edu), y que son los utilizados por Chan en el manual de usuario del programa LOTUS (Chan, 2005). El álgoritmo LOTUS, a grandes razgos, particiona el espacio muestral, y ajusta entonces a cada partición una regresión logística en una o dos variables explicativas. En el caso de este trabajo, solo se revisó el caso para una variable, por lo que es con el que se trabajará. Al hacer ésto, la interpretación se vuelve mucho más sencilla. Al dividir el trabajo entre la estructura del árbol y la regresión logística, Chan y Loh obtienen varias propiedades deseables en cualquier modelo que pretenda ser interpretado y que además explique de manera eficaz el comportamiento de los datos. Algunos puntos importantes son el hecho de que el sesgo a la hora de elegir las variables es negligible, y también que hay un manejo adecuado de datos faltantes (aunque este segundo problema no es concerniente en este momento, es importante notar que el algoritmo lo maneja también).

5.1. Ejemplo 1: Cangrejos

Se ajustará ahora un árbol logístico mediante el algoritmo LOTUS (detallado previamente en el capítulo 3). El problema consiste en clasificar a los cangrejos hembra dentro de dos clases: aquellos con satélites y aquellos sin éstos. Los datos se toman de Ethology (1996) de J. Brockmann obtenidos del libro *An Introduction to Categorical Data Analysis* (Agresti, 1996). Dichos datos muestran posibles factores que pudieran influir en la presencia de cangrejos macho que rodean el nido de una hembra. Se utilizan como variables explicativas a las variables de color, peso, condición de la espina y ancho. De éstas,

se tienen dos categóricas (color y condición de la espina), mientras que las restantes son cualitativas. Recuerdese que en el capítulo 4 se ajustaron estos datos con el algoritmo CART sin mucho éxito. Se verá a continuación como puede corregirse esto utilizando el algoritmo LOTUS. La descripción de las variables, presentada ya en el capítulo 1 se repite a continuación:

- color: color del cangrejo (categorizado como 2, 3, 4 o 5)
- spine: condicion de la espina dorsal (1=dos en buen estado, 2=una es buen estado,3=ninguna en buen estado)
- width: ancho del cangrejo (cm)
- satellite: número de cangrejos satélite
- weight: peso (gramos)
- Y: presencia (1) o ausencia (0) de satélites

El archivo de salida, en el que se presenta el ajuste del modelo LOTUS contiene la siguiente información.

```
Data description file: crabs2.dsc
Learning data file: crabs2.dat
Missing value code: NA

List of variables in data file:
[dependent(d), numerical(s=split only; f=fit only; n=both),
categorical(c=nominal; o=ordinal), excluded(x)]
Column #  Variable name      Variable type
   1      color                    o
   2      spine                    o
   3      width                    f
   4      satellite                x
   5      weight                   f
   6      Y                        d

Summary of variables in data file:
  #column  #n-var  #f-var  #s-var  #c-var  #o-var  #x-var
      6        0       2       0       0       2       1

Number of cases in learning data file = 172
```

```
Number of learning samples (nonmissing responses) = 172
Number of learning samples with one or more missing covariates = 0

Dependent Variable: Y
        Levels     Codes     Count
             0         0        62
             1         1       110

Ordinal Categorical Variables:
            Levels  Categories
color          4    2 3 4 5
spine          3    1 2 3

Model fit: Best simple linear
Deviance estimation: Surrogate model method
P-value for testing best simple model = 0.5000
Minimum node size (MINDAT): 4
Minimum class size in each node (MCLASS): 2
Number of split variable searches: 2
Pruning: Cross-validation
Number of folds for cross-validation = 10
Number of SEs used = 0.00
```

Se observa primero un resumen de las variables: el número de observaciones (tanto las que tienen valores faltantes como las que no). En este caso hay 172 observaciones sin datos faltantes. También se indica que hay 110 categorizadas como 0 y 62 como 1, donde el nivel o la categoría 0 son aquellos cangrejos sin satélites, y el nivel 1 representa a aquellos con satélites.. El modelo se ajustará utilizando regresiones con una sola variable independiente. También se indica el valor-p para hacer pruebas de significancia de los modelos ajustados a cada nodo. El resumen de las variables presenta el número de columna, después el nombre de la variable, y por último el tipo. Las variables explicativas o independientes se separan entre las que se utilizarán solamente como regresoras en el ajuste del modelo (f), las que participan solo para la división de los nodos (s), y las que pueden ser utilizadas en ambos roles(n). Las variables categóricas solo participan en la división, y se separan en dos tipos, ordinales (o) y nominales (c). La diferencia es que en las primeras la categoría está denotada numéricamente, y en las segundas no. Ahora, por otro lado, la variable denotada como (d) es la variable de respuesta, y las variables denotadas como (x) son las que no se utilizarán en el modelo. MINDAT indica el mínimo número de observaciones en los nodos terminales, utilizado durante la construcción del árbol inicial (al que también se ha llamado árbol máximo). En caso de que un nodo tenga el número indicado de observaciones o menos, ya no se dividirá. MCLASS es el menor número de observaciones de cada clase de la variable dependiente

en los nodos. Si en un nodo hay ese número o menos de observaciones, ya no será elegible para dividirlo. Para evitar una posible terminación temprana de la construcción del árbol, se indica que si la variable elegida para hacer la partición del nodo no produce resultados adecuados, se debe intentar particionar con las siguientes dos mejores. Para podar el árbol, el algoritmo permite elegir entre validación cruzada, en la que se indica también el número de subconjuntos que se utilizarán, y la de muestra de prueba. En ambos casos el método utilizado para podar es el utilizado por el algoritmo CART, a saber el método de costo-complejidad. Por último, el número de errores estándar (SE) se utiliza para terminar la poda y seleccionar el tamaño final del árbol. Si se da un valor de cero, se elige el árbol con el menor estimador por validación cruzada de la devianza media (0-SE). Un valor de x arroja el árbol más pequeño cuyo estimador se encuentre a menos de x errores estándar de aquél del árbol 0-SE. Este árbol se llama $x - SE$.

```
Number of terminal nodes in maximal tree = 6
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Pruning Sequence of Nested Subtrees:
 Subtree  Pruned  #Terminal    True           GM
 number    node     nodes      alpha         alpha
    0        -         6     0.000E+00     0.000E+00
    1       16         5     4.422E-03     5.114E-03
    2        8         4     5.914E-03     1.280E-02
    3        4         3     2.770E-02     3.148E-02
    4        2         2     3.577E-02     4.124E-02
    5        1         1     4.754E-02     1.798+308

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Size, CV Mean Deviance and SE of Subtrees:
 Subtree   #Terminal
 number      nodes     CV Mean      CV SE
    0           6     1.189E+00   9.209E-02
    1           5     1.170E+00   9.064E-02
    2           4     1.150E+00   9.021E-02
    3           3     1.176E+00   8.180E-02
    4           2     1.210E+00   7.792E-02
    5           1     1.173E+00   6.871E-02

Subtree 2 is the minimum deviance tree
Subtree 2 is the final optimal tree using SE-rule
```

Las tablas anteriores muestran la sucesión de árboles obtenidos durante la poda y su número de nodos terminales. El árbol máximo es denotado con un 0. En este caso tiene seis nodos terminales. En la primera tabla, la cuarta columna muestra el valor de complejidad α, mientras que la quinta da las medias geométricas de los valores anteriores.

La segunda tabla muestra los estimadores de la devianza media y sus errores estándar correspondientes. En este caso, es posible ver ver que el árbol de devianza media mínima es el número 2, con un valor de 1.150, y por lo tanto es el seleccionado como óptimo. Dicho árbol cuenta con cuatro nodos terminales, y su estructura se presenta a continuación:

```
Structure of Final Tree:
        Total  Cases
  Node  Cases   Fit    Fit_Var    Split_Var     Split      Deviance   Comments
  ================================================================================
    1     172    172  width       color       4.0000E+00  1.9412E+02
    2     150    150  weight      color       3.0000E+00  1.5904E+02
    4     106    106  width       spine       2.0000E+00  1.1350E+02
    8      43     43  weight          <terminal node>     4.6521E+01
    9      63     63  width           <terminal node>     6.2219E+01
    5      44     44  weight          <terminal node>     3.9382E+01
    3      22     22  width           <terminal node>     2.6906E+01

Total deviance of tree = 1.7503E+02
Deviance per observation of tree = 1.0176E+00
Number of terminal nodes of final tree = 4
Total number of nodes of final tree = 7
```

Cada nodo n presenta información sobre el número de observaciones contenidas en éste, la variable con la que se ajusta la regresión logística final en los nodos terminales, la variable utilizada para la división (en el caso de que el nodo no sea terminal) junto con el punto de división elegido, así como la devianza estimada del modelo. La devianza total del árbol es la suma de las devianzas por nodo. La devianza por observación es simplemente la devianza total entre el número de observaciones. En este caso, se tiene una devianza total de 175.03, y la correspondiente devianza por observación de 1.017. El árbol óptimo tiene cuatro nodos terminales y un total de siete nodos.

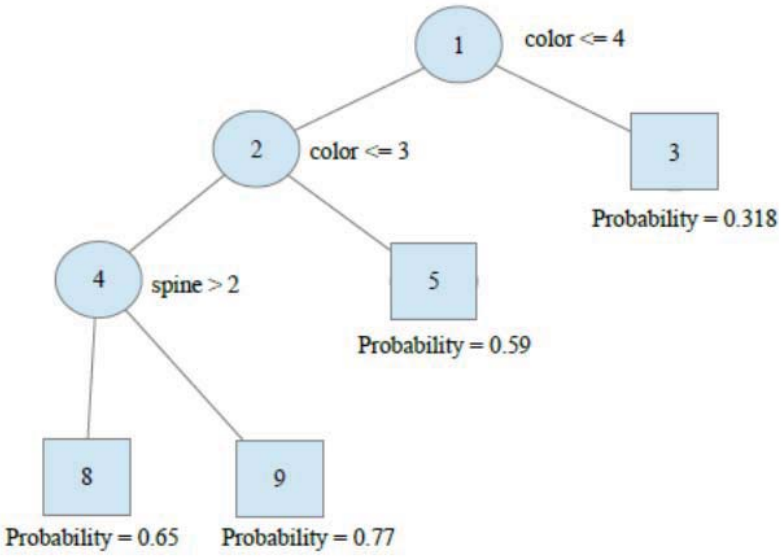
La estructura del árbol también se se puede ver en la siguiente tabla. Los subnodos de un nodo n son $2n$ y $2n + 1$, en el caso de existir.

```
Regression tree:
 ~~~~~~~~~~~~~~~~~
 Node 1: color <= 4.0000E+00
   Node 2: color <= 3.0000E+00
     Node 4: spine <= 2.0000E+00
       Node 8: Probability = 0.6512E+00
     Node 4: spine > 2.0000E+00
```

```
      Node 9: Probability = 0.7778E+00
  Node 2: color > 3.0000E+00
    Node 5: Probability = 0.5909E+00
Node 1: color > 4.0000E+00
  Node 3: Probability = 0.3182E+00
```

Los nodos terminales presentan el resultado del ajuste logístico, que da la probabilidad estimada de pertenecer a la categoría correspondiente. En este caso, en el que solo hay dos clases o categorías, se indica la probabilidad de que una observación el dicho nodo pertenezca a la categoría Y=1, que representa la presencia de cangrejos satélite. Por ejemplo, para el nodo 9, ésta es de 0.77.
Obsérvese una gráfica del árbol anterior:

FIGURA 5.1: Árbol LOTUS: Los nodos intermedios se representan con círculos. La regla para particionar al nodo aparece junto a éste. Si la observación cumple la regla, se va al subnodo izquierdo; en otro caso se va al derecho. Los nodos terminales se representan por rectángulos. Se muestra la probabilidad estimada de que Y=1 por el modelo logístico.



Ahora, para los nodos terminales, se cuenta con información más detallada:

```
Terminal Node Models of Logistic Regression Tree:
 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 Node 8: Deviance = 4.6521E+01
 Total Cases = 43, Cases Fit = 43
 Total Cases with Y=1 = 28
```

```
  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -3.7662E+00  1.7267E+00  -2.1812E+00
  weight        1.8372E-03  7.3769E-04   2.4904E+00
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Node 9: Deviance = 6.2219E+01
Total Cases = 63, Cases Fit = 63
Total Cases with Y=1 = 49

  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -9.2200E+00  5.2288E+00  -1.7633E+00
  width         3.9500E-01  1.9907E-01   1.9843E+00
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Node 5: Deviance = 3.9382E+01
Total Cases = 44, Cases Fit = 44
Total Cases with Y=1 = 26

  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -7.9167E+00  2.5518E+00  -3.1024E+00
  weight        3.7548E-03  1.1759E-03   3.1932E+00
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Node 3: Deviance = 2.6906E+01
Total Cases = 22, Cases Fit = 22
Total Cases with Y=1 = 7

  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -5.8538E+00  6.6939E+00  -8.7450E-01
  width         2.0043E-01  2.6166E-01   7.6599E-01
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Para cada nodo terminal, se tiene el número de observaciones totales y las correspondientes a la categoría 1. Se muestra también el valór estimado de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, junto con sus errores estándar estimados. Los cuatro nodos terminales cuentan con su correspondiente ajuste logístico. Por ejemplo, para un cangrejo perteneciente al nodo tres, la probabilidad de tener satétites es de 0.32. En este caso, la interpretación del resultado es la siguiente: si un cangrejo es del color categorizado como 5, su probabilidad de tener satélites es de 0.32. Para predecir de forma más adecuada la presencia o ausencia de satélites, se usa el modelo simple logístico correspondiente a dicho nodo. En este caso, se tiene que:

$$logit(X) = -5.8538 + 0.20X$$

Donde X corresponde al ancho del cangrejo. Si se conoce el ancho del cangrejo es posible predecir la probabilidad anterior con más precisión.
Análogamente se hace lo mismo para cada nodo terminal. Por ejemplo, una observación cuyo color sea 2 ó 3, y que tenga las dos espinas rotas (categorizado como spine=3) pertenecerá al nodo terminal nueve, en el que también se ajusta un modelo logístico simple con la variable width, el ancho del cangrejo, y que se puede utilizar para predecir la presencia o ausencia de satélites.

Nótese lo siguiente. En el ejemplo anterior se limita a las variables ancho y peso a fungir como regresoras, es decir, no se permite que participen en las divisiones de los nodos. Si se permite que el algoritmo utilice a las dos variables anteriores para generar las particiones del espacio muestral, se obtienen los siguientes resultados:

```
Structure of Final Tree:
        Total  Cases
  Node  Cases   Fit    Fit_Var    Split_Var     Split      Deviance   Comments
  ================================================================================
    1     172    172  width           <terminal node>     1.9412E+02

Total deviance of tree = 1.9412E+02
Deviance per observation of tree = 1.1286E+00
Number of terminal nodes of final tree = 1
Total number of nodes of final tree = 1

Terminal Node Models of Logistic Regression Tree:
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Node 1: Deviance = 1.9412E+02
Total Cases = 172, Cases Fit = 172
Total Cases with Y=1 = 110

  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -1.2253E+01  2.6290E+00  -4.6605E+00
  width         4.9325E-01  1.0177E-01   4.8469E+00
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

En este caso, el mejor modelo encontrado por el algoritmo LOTUS es aquel con un sólo nodo. Es decir, el modelo que mejor explica la presencia o ausencia de satélites es un modelo logístico simple con el ancho del cangrejo como variable explicativa. El modelo queda de la siguiente forma:

$$logit(X) = -12.25 + 0.483X$$

donde X representa el ancho. Si se ve una gráfica de la proporción de cangrejos con satélites según su ancho, queda muy claro el por qué se eligió dicho modelo.

Ancho (cm)	$Y = 0$	$Y = 1$
<23.25	14	5
23.25-24.25	14	4
24.25-25.25	28	17
25.25-26.25	39	21
26.25-27.25	22	15
27.25-28.25	24	20
28.25-29.25	18	15
>29.25	14	14

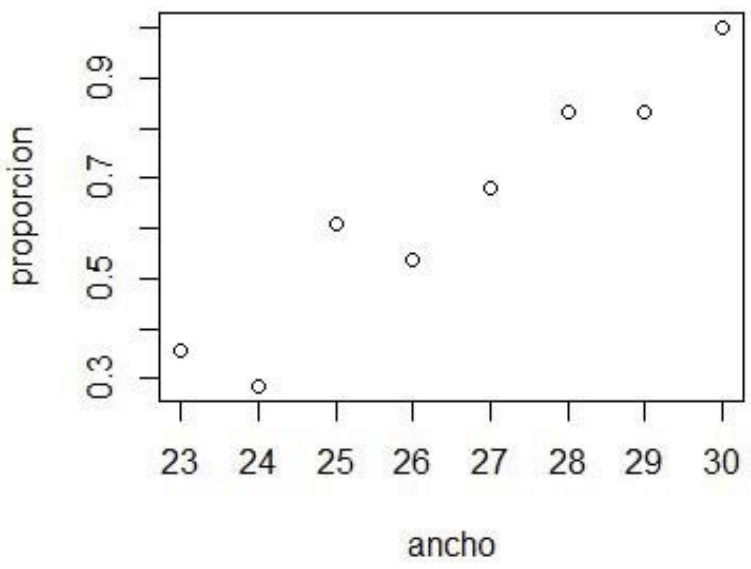


FIGURA 5.2: Gráfica 1: Proporción vs. ancho

Queda claro cómo al ir aumentando el ancho, la proporción de cangrejos con satéites va aumentando también. El modelo anterior explica muy bien el comportamiento de

la variable dependiente en relación con el ancho del cangrejo. De hecho, la devianza de este modelo es de 194, que no es mucho mayor que 175, la devianza del modelo anterior. El hecho de que el algoritmo LOTUS no permita particionar al espacio muestral utilizando las variables numéricas produce este tipo de problemas. De hecho, si se ve un ajuste producido por el algoritmo CART, la primera división corresponde a los siguientes subconjuntos de las observaciones:

$$N_2 = \{x \in N_1 | width < 25.85\}$$

$$N_3 = \{x \in N_1 | width \geq 25.85\}$$

De nuevo N_1 denota al nodo raíz y N_2 y N_3 a sus subnodos, o nodos hijos. En este caso, de las 94 observaciones en N_3, sólo 18 pertenecen a la categoría de cangrejos sin satélites, mientras que las restantes 76 sí tienen satélites. Si se ajusta una regresión logística tomando solamente a las observaciones con un ancho mayor a 25.85, se obtiene el modelo siguiente:

$$logit(X) = -12.12 + 0.49X$$

en el que la devianza residual es de 86.6 y la nula de 91.8, con 92 y 93 grados de libertad respectivamente. En el caso del modelo con todas las observaciones, estos valores son: devianza nula de 224.87 con 171 grados de libertad y devianza residual de 194.12 con 170 grados de libertad.

Se puede entonces calcular la estadística G_1 y G_2 para ambos subconjuntos de observaciones:

$$G_1 = 91.8 - 86.6 = 5.2$$

$$G_2 = 224.9 - 194.1 = 10.8$$

Recuérdese que bajo la hipótesis de que $\beta_1 = 0$ la estadística G se distribuye χ^2 con un grado de libertad. Se obtienen los valores-p para ambos modelos: 0.022 para el primero y 0.001 para el segundo. Ahora, si se toman las observaciones con un ancho menor a 25.85, obtenemos un modelo logístico con devianza nula de 106.85 con 77 grados de libertad y devianza residual 105.71 con 76 grados de libertad. Esto arroja un valor de $G = 1.14$, y un valor-p de 0.28. Mientras que para el modelo con todas las observaciones la inclusión de la variable ancho es significativa a un 99 %, para los otros dos ya no. Esto dice que la incorporación de la variable ancho, una vez que particionado el espacio muestral, ya no es significativa. Entonces, si se utiliza el ancho para dividir primero al conjunto de observaciones, esta variable pierde mucha relevancia después. Al limitarse al uso de las variables numéricas para hacer el ajuste logístico, el algoritmo LOTUS pierde mucha potencia, pues en casos como éste, usar la variable numérica para particionar es mucho más eficiente que usarla solamente como regresora.

5.2. Ejemplo 2: Producción de automóviles en Estados Unidos

Los datos con los que se trabaja en este ejemplo presentan información sobre la producción de coches en EUA. La descripción de las variables es la siguiente:

Variable	mpg	displacement	hp	weight	acc
min.	9	68	46	1613	8
1st quantile	17.5	105	75.75	2226	13.7
median	23	151	95	2822	15.5
mean	23.51	194.8	105.08	2979	15.52
3rd quantile	29	302	130	3618	17.18
max.	46.6	455	230	5140	24.8
NA's	8	/	6	/	/

```
List of variables in data file:
 [dependent(d), numerical(s=split only; f=fit only; n=both),
 categorical(c=nominal; o=ordinal), excluded(x)]
 Column #  Variable name      Variable type
    1      car                      d
    2      milesperga               n
    3      cylinder                 o
    4      displaceme               n
    5      horsepower               n
    6      weight                   n
    7      accelerati               n
    8      year                     c
```

```
   9      origin                  x

Summary of variables in data file:
  #column  #n-var  #f-var  #s-var  #c-var  #o-var  #x-var
      9        5       0       0       1       1       1

Number of cases in learning data file = 406
Number of learning samples (nonmissing responses) = 406
Number of learning samples with one or more missing covariates = 14

Dependent Variable: car
        Levels     Codes     Count
       NON-USA         0       152
           USA         1       254

Ordinal Categorical Variables:
            Levels  Categories
cylinder       5    3 4 5 6 8

Nominal Categorical Variables:
            Levels  Categories
year          13    70 71 72 73 74
                    75 76 77 78 79
                    80 81 82
```

Se tiene información sobre el año de producción, la aceleración, el peso, caballos de fuerza, la cilindrada, el número de cilindros, y las millas por galón. La variable de respuesta (car) indica si el auto fue producido fuera de EUA (0), o en EUA (1). El archivo de nuevo presenta el número de variables de cada tipo, así como las categorías de las variables ordinales y nominales. En este ejemplo, se trabaja con siete variables de respuesta, de las cuales cinco son numéricas y se les permite participar en las divisiones y en el ajuste; y dos son categóricas, una nominal y una ordinal.

```
Model fit: Best simple linear
Deviance estimation: Surrogate model method
P-value for testing best simple model = 0.5000
Minimum node size (MINDAT): 4
Minimum class size in each node (MCLASS): 3
Number of split variable searches: 2
Pruning: Cross-validation
Number of folds for cross-validation = 10
Number of SEs used = 0.00

Number of terminal nodes in maximal tree = 12
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Pruning Sequence of Nested Subtrees:
 Subtree  Pruned  #Terminal    True           GM
 number    node     nodes      alpha         alpha
    0       -        12     0.000E+00     0.000E+00
    1      77        11     0.000E+00     0.000E+00
    2       5         9     0.000E+00     0.000E+00
    3      38         8     9.452E-03     1.036E-02
    4      19         7     1.135E-02     1.656E-02
    5       9         5     2.418E-02     2.559E-02
    6       4         3     2.708E-02     4.589E-02
    7       1         1     7.776E-02     1.798+308

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Size, CV Mean Deviance and SE of Subtrees:
 Subtree   #Terminal
 number      nodes     CV Mean      CV SE
    0          12     7.690E-01   1.154E-01
    1          11     6.663E-01   1.012E-01
    2           9     6.663E-01   1.012E-01
    3           8     6.703E-01   1.009E-01
    4           7     6.633E-01   1.000E-01
    5           5     6.553E-01   9.913E-02
    6           3     5.801E-01   7.486E-02
    7           1     6.431E-01   4.721E-02

Subtree 6 is the minimum deviance tree
Subtree 6 is the final optimal tree using SE-rule
```

Se continúa revisando el archivo de salida. En este caso, se indica que se eligió ajustar modelos logísticos con una sola variable independiente. También se indica que el mínimo número de datos en un nodo antes de dejar de dividirlo es de 4, y el de clases es 3. El método para podar el árbol elegido fué por validación cruzada. Los resultados obtenidos muestran que el árbol máximo cuenta con 12 nodos terminales. El proceso de poda por validación cruzada generó ocho árboles, partiendo del máximo y llegando al mínimo, que sólo cuenta con el nodo raíz. En la segunda tabla se encuentran los estimadores de la devianza media para cada árbol. Estos valores son una medida de qué tan bueno es el ajuste para cada árbol. En este caso, el subárbol 6 es el de menor devianza media. El programa LOTUS permite elegir el número de errores estándar de distancia entre el árbol de devainza estimada mínima y el árbol seleccionado, es decir el árbol óptimo. En este caso se usa 0, por lo que el árbol de mínima devianza y el árbol óptimo son el mismo. En caso de elegir por ejemplo un valor de 1 error estándar, se buscaría el mejor entre todos los árboles con error estándar $5.801E-01 \pm 1$.

Una vez elegido el árbol óptimo, se presenta más detalladamente su estructura.

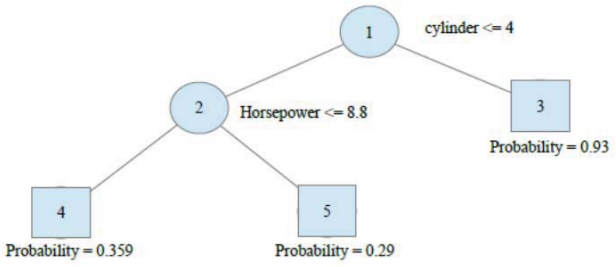
```
Structure of Final Tree:
        Total  Cases
  Node  Cases   Fit    Fit_Var    Split_Var     Split      Deviance   Comments
  ================================================================================
    1     406    406  displaceme  cylinder    4.0000E+00  2.5934E+02
    2     211    211  displaceme  horsepower  8.8000E+01  2.1504E+02
    4     156    156  displaceme      <terminal node>     1.5712E+02
    5      55     55  displaceme      <terminal node>     2.0214E+01
    3     195    195  displaceme      <terminal node>     1.8866E+01

Total deviance of tree = 1.9620E+02
Deviance per observation of tree = 4.8326E-01
Number of terminal nodes of final tree = 3
Total number of nodes of final tree = 5

Regression tree:
~~~~~~~~~~~~~~~~
Node 1: cylinder <= 4.0000E+00
  Node 2: horsepower <= 8.8000E+01
    Node 4: Probability = 0.3590E+00
  Node 2: horsepower > 8.8000E+01
    Node 5: Probability = 0.2909E+00
Node 1: cylinder > 4.0000E+00
  Node 3: Probability = 0.9333E+00
```

La gráfica obtenida es la siguiente:

FIGURA 5.3: Árbol LOTUS: Los nodos intermedios se representan con círculos. La regla para particionar al nodo aparece junto a éste. Si la observación cumple la regla, se va al subnodo izquierdo; en otro caso se va al derecho. Los nodos terminales se representan por rectángulos. Se muestra la probabilidad estimada de que Y=1 por el modelo logístico.



El subárbol final tiene 5 nodos, de los cuales tres son terminales. El programa muestra el número de observaciones en cada uno de los nodos, así como la variable con la que se ajusta la regresión logística, la variable con la que se particionó el nodo, y la devianza estimada en cada nodo. Por ejemplo, el nodo raíz (1) contiene las 406 observaciones iniciales. La variable con la que se obtuvo el ajuste logístico simple de menor devianza fue displacement, que indica la cilindrada del auto. El nodo se dividió de acuerdo a la siguiente regla:

$$N_2 = \{x \in N_2 | \text{ cylinder} \leq 4.0000E+00\}$$

$$N_3 = \{x \in N_1 | \text{ cylinder } 4.0000E+00\}$$

Donde se denota como N_1 al nodo raíz y como N_2 y N_3 a sus nodos hijos. De éstos, el nodo N_3 es terminal. Su ajuste logístico también se realizó con la variable displacement, e indica que si una observación está en dicho nodo, la probabilidad de que el auto haya sido producido en Estados Unidos es de 0.933.
Para ver con más detalle la estructura de los nodos terminales, el programa presenta también la siguiente tabla:

```
Terminal Node Models of Logistic Regression Tree:
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Node 4: Deviance = 1.5712E+02
Total Cases = 156, Cases Fit = 156
Total Cases with Y=1 = 56

  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -7.7416E+00  1.2751E+00  -6.0714E+00
  displaceme    6.7781E-02  1.1920E-02   5.6863E+00
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Node 5: Deviance = 2.0214E+01
Total Cases = 55, Cases Fit = 55
Total Cases with Y=1 = 16

  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -3.2570E+01  9.8991E+00  -3.2902E+00
  displaceme    2.3760E-01  7.2526E-02   3.2761E+00
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Node 3: Deviance = 1.8866E+01
Total Cases = 195, Cases Fit = 195
Total Cases with Y=1 = 182
```

```
  Variable     Coefficient   Std Error    T-Value
  ------------------------------------------------
  Intercept    -2.4093E+01  8.6006E+00  -2.8014E+00
  displaceme    1.4152E-01  5.0341E-02   2.8112E+00
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Se muestra en la tabla anterior el modelo logístico simple para cada nodo. En el caso del nodo N_3, se tienen 195 observaciones, de las cuales 182 pertenecen a autos producidos en EUA, lo que es coherente con el resultado anteriormente mencionado. Los coeficientes estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ también son mostrados junto con sus errores estándar. Esto permite, a diferencia del algoritmo CART, ya no sólo categorizar cada nodo terminal según una clase, sino tener una idea de cómo están relacionadas la variable dependiente y la variable explicativa, para la partición de la muestra que pertenece a cada nodo. En este caso, para el nodo 3, una posible interpretación sería la siguiente: la probabilidad de que un auto con más de cuatro cilindros haya sido producido en EUA es de 0.93. Para este subconjunto de las observaciones, la variable numérica independiente que mejor explica el comportamiento de la variable de respuesta es displacement, para la que se estima el logit como sigue:

$$logit(X) = -24.093 + 0.1415X$$

donde X representa a la variable *displacement*. A partir de este ajuste, se puede predecir cuando un auto con más de 4 cilindros fué producido en Estados Unidos. De la misma forma, se puede obtener la probabilidad para las observaciones pertenecientes a cada nodo terminal utilizando el ajuste logístico correspondiente, y recordando que las observaciones pertenecientes a cada uno de estos nodos son aquellas que cumplen (o no) con las reglas mediante las cuales se particionó el espacio muestral.

Los ejemplos anteriores muestran dos posibles implementaciones del algoritmo LOTUS. La potencia de dichos modelos reside en que se combina el uso de árboles de clasificación con la regresión logística, lo que permite por un lado comprender de forma clara la relación entre las variables independientes y la variable de respuesta, además de que, al no solamente asignar una categoría a los nodos terminales sino ajustar una regresión, permite hacer una lectura más robusta de las observaciones.

Capítulo 6

Conclusiones

Todo modelo estadístico no es más que una aproximación para intentar explicar el comportamiento de cierta variable que representa a algún objeto de la realidad, por lo que nunca va a poder llegar a ser perfecto. La regresión logística, como tal, tiene sus ventajas y desventajas. Se sabe que para poder explicar y predecir de forma correcta el comportamiento presente y futuro de una variable se requeriría de una cantidad enorme de variables explicativas. Lo que se hace al utilizar una regresión es tomar a aquellas variables que se relacionan de forma más clara y precisa, y que explican de forma significativa el comportamiento de la variable respuesta. Evidentemente, mientras más variables se incluyan en el modelo, la precisión en la predicción irá aumentando, pero también la complejidad de éste. Por un lado, se tendrá un modelo más correcto -puesto que la relación modelada se irá acercando a la real-, pero por otro lado la interpretación será también cada vez más difícil de hacer. Así, en un modelo logístico con muchas variables independientes se corre el riesgo de que los coeficientes sean cada vez más difíciles de interpretar, lo que no es deseado, puesto que a pesar de que tal vez el modelo sea más preciso, sin la posibilidad de interpretarlo se pierde una de las funciónes principales por las que se está utilizando la regresión logística.

Se revisáron también en éste trabajo los árboles de clasificación y regresión, una herramienta relativamente nueva, que se ha encontrado también es útil para para clasificar datos dentro de algúna categoría, así como para explicar la relación entre variables (igual que hace la regresión logística). El algorítmo revisado en éste trabajo, llamado CART, desarrollado por Briemann, Freidman, Olshen y Stone, es uno de los primeros algorítmos en utilizar la estructura de árbol con objetivos estadísticos. En tanto que modelo estadístico, los árboles tienen también sus ventajas y desventajas: por un lado, la complejidad que se presenta en regresión logística cuando existen muchas variables independientes se modela mediante la estructura del árbol, con lo que se muestran -al ir particionando el espacio muestral- de forma clara las condiciones que debe cumplir una

observiación para pertenecer a cierta categoría. Por otro lado, al tratar de arrojar un modelo sencillo, se pierde mucha de la fuerza de predicción, y se puede llegar incluso a sobre-simplificar el modelo. Otro problema es que si se trata de igualar la fuerza predictiva de otros modelos, se obtienen árboles muy grandes, con los que de nuevo se pierde el objetivo de lograr modelos de fácil interpretación.

Para lograr evitar los problemas de los modelos anteriores, parece que una solución natural es combinar la sencillez interpretativa de los árboles con la precisión de la regresión logística. Resulta que no es tan sencillo como parece combinar el uso de éstos modelos. Se debe poner especial atención a como se realizan las particiones del espacio muestral, así como a la forma de elegir los modelos que se ajustarán a cada nodo. La solución que se revisa en este trabajo es la propuesta por Chan y Loh: el algorítmo LOTUS. éste modelo tiene varias ventajas. Por ejemplo, se evita que haya sesgo al particionar el espacio muestral seleccionando de forma adecuada a las variables. Además, permite ajustar a cada nodo regresiónes logísticas, con lo que en vez de simplemente asignar una categoría a éstos, se les asigna un modelo sencillo que muestra la relación entre una de las variables independientes y la variable respuesta. Al combinar la regresión logística con los árboles de clasificación, se obtiene un árbol sencillo de interpretar pero que no sacrifica la capacidad predictiva de los modelos logísticos.

Debido al potencial de los modelos logísticos, y a su aplicación en diversas áreas de la ciencia, una propiedad que se ha vuelto necesaria es la de su interpretación. El problema es que al buscar un modelo con una interpretación sencilla se corre el riesgo de sobresimplificarlo, lo que lleva a un modelo cuyas predicciones pueden no ser precisas. Idealmente, se busca balancear el ajuste del modelo con la complejidad.

El modelo CART utiliza árboles binarios para tratar de modelar problemas de clasificación y de regresión lineal de una forma clara y natural, dando un entendimiento de la estructura subyacente al problema. Por otro lado, el algoritmo LOTUS, aplica la estructura de los árboles binarios para modelar problemas logísticos. La idea detrás de un acercamiento gráfico (es decir mediante árboles) es la de «divide y vencerás». Al particionar el espacio muestral de forma que en cada partición se ajuste una regresión logística simple, resulta en un modelo que combina las bondades de la interpretación gráfica con el potencial predictivo de la regresión logística. De esta forma se obtiene un modelo en el que se facilita la interpretación sin perder la capacidad predictora.

Es posible decir que los modelos CART son mejores que los modelos logísticos, o que los modelos del algoritmo LOTUS lo son? En Estadística y en general cuando se deben analizar datos, es dificil decidir cuándo un modelo es mejor que otro, y más aún comprender el porque. Es por ésto que podemos concluir que la única forma de obtener el mayor provecho es no limitarse al uso de un solo modelo. Como suele suceder con las herramientas estadíasticas, el combinarlos y así complementarlos es una de las mejores formas de obtener mejores resultados.

Capítulo 7

Bibliografía

Hosmer, David & Lemeshow, Stanley. 2000. Applied Logistic Regression -second edition. Wiley. EUA. 375 p.

Agresti, Alan. 2002. Categorical Data Analysis. Wiley. EUA. 710 p.

Breiman, Leo, et.al. 1984. Classification and Regression Trees. Chapman & Hall/CRC. EUA. 358 p.

Loh, Wei-Yin & Chan, Kin-Yee. 2004. LOTUS: an Algorithim for Building Accurate and Comprehensible Logistic Regression Trees.Journal of Computational and Graphical Statistics, 13(4):826-852.

Kleinbaum, David & Klein, Mitchel. 1994. Logistic Regression: A Self-learning Text second edition. Springer. EUA. 513 p.

Cramer, J.S. 2003. The origins and development of the logit model.

Chan, Kin-Yee. 2005. LOTUS User Manual.

Loh, Wei-Yin. Logistic Regression Tree Analysis, in Handbook of Engineering Statistics, H. Pham, ed., 537549, Springer, 2006.

Landwehr, Niels. 2003. Logistic Model Trees. University of Freiburg. Germany. 92 p.

Montgomery, Douglas C. 2001. Introduction to Linear Regression Analysis third edition. Wiley. EUA. 641 p. R Data Analysis Examples: Logit Regression. UCLA: Institute for Digital Research and Education. http://www.ats.ucla.edu/stat/r/dae/logit.htm

Casella, George & Berger, Roger. 2002. Statistical Inference, second edition. Duxbury. EUA. 660 p.

Draper, Norman R. & Smith, Harry. 1966. Applied Regression Analysis. Wiley. EUA. 709 p.

Demaris, Alfred. 1992. Logit Modeling. Practical Applications. Sage University Papers. Sage Publications. EUA. 87 p.

Menard, Scott. 1995. Applied Logistic Regression Analysis. Sage University Papers. Sage Publications. EUA. 96 p.

Pham Haang. 2006. "Logistic Regression Tree Analysis", en Springer Handbook of Engineering Statistics. Springer. EUA. p. 537-549.