Bioestadística (Master en Bioinformática)

Bloque: AMECP

Análisis de Modelos Estadísticos de Comparación y Predicción

Sesión: Regresión logística y Análisis de curvas ROC

Manuel Franco
Dpto. Estadística e Investigación Operativa

Índice

1. Introducción

El Análisis de Regresión Logística es una técnica estadística que se encarga de establecer la relación entre un conjunto de variables disponibles en nuestro campo de estudio experimental (a través de sus observaciones o datos registrados), analizando la relevancia de cada una de ellas en la relación, con el objetivo de estimar o predecir una variable respuesta o dependiente de tipo cualitativo (no continua) y proporcionar un mecanismo de discriminación en la población.

En nuestro caso, se utilizará para clasificar los individuos de la población bajo estudio mediante un modelo estadístico de las variables o factores incluidos en el modelo (covariables) que permite la predicción o diagnóstico de un suceso de interés clínico determinado por la variable respuesta, y que a su vez contempla la importancia de cada covariable en la relación con la respuesta.

Por un lado, podría plantearse un modelo de regresión lineal para abordar dicho objetivo. No obstante, como hemos visto, la regresión lineal permite obtener el modelo que mejor se ajusta a la información disponible entre las variables para estimar la media de nuestra variable respuesta o bien predecir el valor de ésta. En este sentido, el tipo de característica de interés clínico que se pretende analizar (variable respuesta) debe ser tenido en cuenta a la hora de elegir la técnica estadística más adecuada en cada situación.

Por ejemplo, en la mayoría de las tareas clínicas se presenta la disyuntiva entre dos únicos casos para el diagnóstico de una enfermedad, el paciente o individuo padece la enfermedad (positivo) o no

la padece (negativo), es decir, la característica de interés (variable respuesta) es de tipo dicotómico, esto es, sólo tiene dos valores y en general se representan por 0 y 1. En este caso, el objetivo del análisis, más que en el propio valor de la variable Y, se centra en su probabilidad, es decir, en predecir la probabilidad de que ocurra el suceso objeto de estudio dado por $Y = 1$ a través de la información suministrada por las variables o factores clínicamente relevantes para realizar dicha tarea. Este suceso $Y = 1$ puede representar cualquier situación de interés clínico, desde padecer una infección o morir, hasta recuperarse o sobrevivir a una enfermedad.

En este sentido, aunque pueda plantearse un modelo de regresión lineal, el objetivo que se persigue pone de manifiesto ciertas limitaciones e incoherencias del modelo lineal, que se resuelven fácilmente teniendo en cuenta el carácter de la variable respuesta, siendo una de las técnicas adecuadas el modelo de regresión logístico. Así, por ejemplo, si planteáramos un modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 x + \varepsilon \;\Leftrightarrow\; E(Y|x) = \beta_0 + \beta_1 x,$$

siendo $Y = 1$ la presencia del suceso e $Y = 0$ la ausencia del mismo en las condiciones x, por lo que $E(Y|x) = P(Y = 1|x)$, i.e.,

$$p_x = P(Y = 1|x) = \beta_0 + \beta_1 x.$$

En consecuencia, el modelo de regresión lineal estimado $\widehat{y}_x = \widehat{p}_x$, predecirá la probabilidad de pertenecer al grupo de la población que presenta el suceso de interés clínico cuando se dan las características x. Esto supone uno de sus primeros inconvenientes, dado que $p_x \in [0, 1]$ y el modelo lineal $\beta_0 + \beta_1 x$ no garantiza valores factibles en $[0, 1]$, y además, la falta de normalidad (Y es discreta) o de heterogeneidad de varianzas produciría un resultado de clasificación de la población no necesariamente óptimo y de difícil interpretación en términos de probabilidad de pertenencia o no de un individuo al grupo del suceso en estudio.

El modelo de regresión logístico proporciona una forma de resolver estos inconvenientes del modelo de regresión lineal, introduciendo una **función de enlace** entre el modelo lineal y la probabilidad del suceso de interés dado por $Y = 1$, que permite discriminar entre los grupos de la población mediante un modelo utilizado para predecir la presencia del suceso en los individuos y cuyos términos representan la importancia de las variables o factores en dicho modelo.

2. Modelo de regresión logística

Una función de enlace entre la probabilidad de ocurrencia del suceso y el modelo lineal debe garantizar que sus valores sean factibles, es decir, que proporcione valores estimados o pronosticados en $[0, 1]$ para dicha probabilidad. Obsérvese que hay varias funciones de enlace que se utilizan para este propósito, por ejemplo los modelos de probabilidad están definidos por una función de distribución con ciertas características de monotonía, continuidad y convergencia; en particular garantiza todos sus valores en el intervalo $[0, 1]$, y por tanto puede ser utilizada como función de enlace.

Modelo logit simple

En nuestro caso, nos centraremos en la función de distribución logística, una de las funciones de enlace más habituales para establecer el modelo de regresión logístico por la adecuada interpretación de sus términos y su interés en las áreas de Ciencias de la Salud, de esta distribución proviene el nombre de esta técnica estadística, modelo logit, y a partir de una covariable o factor, el modelo de regresión logístico con función de enlace la distribución logística, $F(x) = \frac{e^x}{1+e^x}$ con $x \in \mathbb{R}$, viene dado por:

$$p_x = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

o equivalentemente, a través del concepto de *odds* del suceso

$$odds_x = \frac{p_x}{1-p_x} = \exp(\beta_0 + \beta_1 x) \;\Leftrightarrow\; \log odds_x = \log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x$$

conocido éste último, como modelo logit en el análisis de regresión logística, y representa cuanto mayor es la probabilidad de pertenecer al grupo de la población que presenta el suceso de interés clínico $Y = 1$ frente a que no pertenezca a dicho grupo $Y = 0$ cuando se dan las características x, ya que

$$odds_x = \frac{p_x}{1-p_x} = \frac{P(Y=1|x)}{P(Y=0|x)}.$$

Modelo logit múltiple

En general, para llevar a cabo la clasificación de los individuos de la población bajo estudio según su pertenencia o no al suceso de interés clínico $Y = 1$, el análisis de regresión logística proporciona el modelo para predecir su probabilidad de pertenencia a través de la información suministrada por las variables o factores relevantes, y por consiguiente, el modelo de regresión logístico se extiende al caso múltiple para incluir todas las componentes $\mathbf{x} = (x_1, x_2, ..., x_k)$ relevantes en el estudio del suceso $Y = 1$, como sigue:

$$\log odds_{\mathbf{x}} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \;\Leftrightarrow\; p_{\mathbf{x}} = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}. \tag{2.1}$$

2.1. Medidas de riesgo y *odds-ratio*s

Para un mejor entendimiento del modelo de regresión logístico, introducimos primero el concepto de *odds-ratio*, mostrando su relación con el riesgo relativo a través de la noción de *odds* que compara las probabilidades de ocurrencia y no ocurrencia. Y posteriormente, la relación entre *odds-ratio* y los coeficientes del modelo.

Riesgo y *odds*

En este sentido, el *odds* de un suceso está determinado por su probabilidad de ocurrencia, y viceversa, esta probabilidad está determinada por su *odds*. Por ejemplo, en un estudio de supervivencia y factores de riesgo de muerte en pacientes ingresados en la Unidad de Cuidados Intensivos del Baystate Medical Center de Springfield en Massachusetts (Hosmer and Lemeshow, 2000), se estima que el 20 % de los pacientes que ingresan con un determinado diagnóstico clínico fallecen (esto es que cada paciente tiene una probabilidad 0.2 de no sobrevivir a su enfermedad), y por consiguiente, el 80 % de ellos superan su enfermedad, esto significa que un paciente que ingresa en la UCI por diagnóstico de enfermedad con riesgo de muerte es 4 veces menos probable de que fallezca frente a que no fallezca,

$$odds(morir) = \frac{P(morir)}{1 - P(morir)} = \frac{0.20}{0.80} = \frac{1}{4}$$

o equivalentemente,

$$P(morir) = \frac{odds(morir)}{1 + odds(morir)} = \frac{1/4}{1 + 1/4} = 0.2;$$

recíprocamente, un paciente que ingresa en la UCI por diagnóstico de enfermedad con riesgo de muerte es 4 veces más probable de que sobreviva frente a que no sobreviva,

$$odds(sobrevivir) = \frac{0.80}{0.20} = 4 \;\Leftrightarrow\; P(sobrevivir) = \frac{4}{1+4} = 0.8.$$

Como vemos, hay una relación directa entre el *odds* de un suceso y su probabilidad. Si aumenta la probabilidad aumenta su *odds*, y si disminuye la probabilidad de ocurrencia disminuye su *odds*. No obstante, el cambio de tendencia en probabilidad se centra en el 0.5 (el 50 % de los pacientes se mueren y el 50 % no se mueren), lo que indica que es un suceso totalmente aleatorio (como lanzar una moneda). En este caso, sería igualmente probable que un paciente ingresado en la UCI se muriera frente a que no se muriera, siendo entonces 1 el valor del cambio de tendencia en el *odds*.

Riesgo relativo y *odds-ratio*

En este contexto, la probabilidad de ocurrencia del suceso (un paciente ingresado en la UCI fallece, $Y = 1$) o en términos porcentuales, puede estar afectada por la presencia o no de algún factor de riesgo F clínicamente relevante para el diagnóstico de la enfermedad, así como en el estudio de su desarrollo y evolución. La influencia de este factor de riesgo introduce un concepto de gran interés en el campo de las Ciencias de la Salud, el riesgo relativo de ocurrencia del suceso en presencia del factor respecto de su ausencia:

$$\text{Riesgo relativo de morir en presencia de F respecto a no F} = \frac{P(morir|F)}{P(morir|\text{no } F)},$$

y en particular, de actualidad en la investigación epidemiológica para determinar grupos de riesgo en la transmisión o contagio, es decir, el riesgo relativo a enfermar cuando se está en un grupo A con respecto a otro grupo B:

$$\text{Riesgo relativo de enfermar en A respecto a B } = \frac{P(enfermar|A)}{P(enfermar|B)}.$$

Este concepto, en general, representa el riesgo de ocurrir un suceso relativo a dos grupos o niveles de un factor, siendo de especial interés cuando se analiza el haber estado expuesto a determinadas situaciones de riesgo o la presencia de ciertas condiciones frente al estado contrario. Por ejemplo, estar vacunado o no, ser diabético o no, ser fumador o no, ser hipertenso o no, ...

A partir del *odds* del suceso en cada grupo, $odds_A$ y $odds_B$, por similitud al riesgo relativo, el *odds-ratio* está dado por la tasa de *odds*:

$$odds\text{-}ratio_{(A,B)} = \frac{odds_A}{odds_B},$$

y representa cuanto más probable es que ocurra el suceso (enfermar) frente a que no ocurra en A con respecto a B.

Observar que estos dos últimos, riesgo relativo y *odds-ratio*, coinciden en el valor 1 como punto del cambio de tendencia, dado que tener la misma probabilidad de enfermar en A y en B (riesgo relativo=1) es equivalente a la igualdad de las cantidades de cuanto más probable es enfermar frente a no enfermar en cada grupo (*odds-ratio*=1), aunque esta equivalencia no se mantiene entre el resto de valores del riesgo relativo y *odds-ratio*.

Coeficientes y *odds-ratio*

A través de estos conceptos, podemos volver a la expresión (2.1) del modelo logit, e interpretar los coeficientes del modelo de regresión logístico como cuantificadores de la importancia de las variables o factores. Por ejemplo, considerando que se mantienen constantes todas las características $\mathbf{x}$ excepto en la componente x_j que se modifica en una unidad, el $odds\text{-}ratio_{x_j}$ dado por la tasa entre $odds_{(x_1,\ldots,x_{j-1},x_j+1,x_{j+1},\ldots,x_k)}$ y $odds_{\mathbf{x}}$ representa cuanto más probable es que ocurra $Y = 1$ frente a que

no ocurra cuando se aumenta una unidad en x_j manteniendo constantes el resto de componentes de $\mathbf{x}$, es decir,

$$\textit{odds-ratio}_{x_j} = \frac{odds_{(x_1,\ldots,x_{j-1},x_j+1,x_{j+1},\ldots,x_k)}}{odds_{\mathbf{x}}} = \exp(\beta_j)$$

y por tanto, indica cuanto se modifica la ocurrencia del suceso frente a su ausencia cuando la variable x_j aumenta una unidad.

En este sentido, para una correcta interpretación y cuantificación de la importancia de los términos del modelo de regresión logístico, debemos distinguir entre el tipo de variables y factores que forman parte del mismo:

- Variable cuantitativa: El coeficiente mide en cuanto aumenta o disminuye el logaritmo del *odds-ratio* al aumentar la variable en una unidad.
- Variable categórica binaria: El coeficiente mide en cuanto aumenta o disminuye el logaritmo del *odds-ratio* bajo la exposición o presencia del factor de riesgo respecto a su ausencia. Si es negativo se dice que se trata de un factor de protección.
- Variable categórica nominal: La variable sólo identifica las categorías, no valores de estas. Para contemplar esta variable en el modelo debe utilizarse variables ficticias indicadoras de cada categoría, lo que nos permite interpretar cada coeficiente como el aumento o disminución del logaritmo del *odds-ratio* de esta categoría con respecto a una categoría base de referencia o de control. Observar que corresponden a indicadoras de cada categoría de la variable nominal, por lo que se incluirán o no en bloque en el modelo logit.
- Variable categórica ordinal: La variable identifica en una escala ordenada las categorías. Asumiendo el mismo grado o cuantificación entre categorías contiguas, algunos expertos consideran que puede interpretarse de modo similar a una variable cuantitativa, el coeficiente mide en cuanto aumenta o disminuye el logaritmo del *odds-ratio* al pasar de un nivel al siguiente en la escala ordenada. Sin embargo, otros autores descartan esta opción, señalando que debe procederse al igual que una variable categórica nominal.

No obstante, hay que tener en cuenta las limitaciones del tamaño muestral a la hora de incluir las covariables (variables o factores) en el modelo, y el desglose correspondiente de estas covariables cuando son de tipo categórico. Algunos expertos recomiendan un mínimo de 10 individuos observados en el suceso de menor representación de la variable respuesta por cada covariable, o al menos en los cruces de cada covariable con la respuesta.

2.2. Estimación y análisis del modelo logístico

Una vez planteado el modelo de regresión logístico (2.1) para predecir la probabilidad de ocurrencia del suceso de interés clínico y para clasificar los individuos de una población según su pertenencia o no a este grupo objeto de estudio, se procede a la estimación de los coeficientes $\widehat{\beta}_j$ de las covariables a partir de las observaciones registradas en una muestra de individuos de la población, utilizando las variables o factores de riesgo relevantes, incluidas las componentes ficticias de las variables categóricas si fuese necesario. Esto nos proporcionará un modelo estimado o ajustado, y por consiguiente, los valores pronosticados $\widehat{p}_{\mathbf{x}}$ para las probabilidades del suceso $Y = 1$ en cada una de las características o condiciones determinadas por $\mathbf{x}$.

Estimación del modelo logístico

Para llevar a cabo dicha estimación del modelo, se aplica el método de estimación de máxima verosimilitud para los coeficientes del modelo de regresión logística, es decir, calcular los valores para los coeficientes en los que alcanza su máximo valor el logaritmo de la función de verosimilitud (probabilidad conjunta de la muestra registrada de tamaño n), de forma abreviada:

$$\log L(y_1, ..., y_n) = \sum_{i=1}^{n} \left(y_i \log p_i + (1 - y_i) \log(1 - p_i)\right),$$

siendo necesario el uso de técnicas de optimización numérica para aproximar las soluciones mediante algoritmos iterativos, así como para los cálculos de sus errores estándar, lo que da lugar al modelo de regresión logístico ajustado:

$$\widehat{p}_{\mathbf{x}} = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k)}.$$

Además, observar que este método equivale a minimizar la **deviance**, dada por la suma de n términos que miden la desviación del modelo a los datos:

$$D = \sum_{i=1}^{n} d_i^2 = -2 \log L(y_1, ..., y_n),$$

donde cada d_i representa la desviación en la observación i.

Cabe señalar que la proximidad a 1 de la verosimilitud a partir de los valores pronosticados expresa la eficiencia de este modelo para representar la realidad, o equivalentemente, la proximidad a 0 del logaritmo de la verosimilitud o de la deviance.

Análisis del modelo de regresión logística

En la presentación de los resultados de un análisis de regresión logística, es habitual incluir junto a las estimaciones de los coeficientes $\widehat{\beta}_j$ de cada variable x_j, la estimación de su error estándar, el contraste de significación de Wald, su *odds-ratio* estimada $\widehat{\textit{odds-ratio}}_{x_j} = \exp \widehat{\beta}_j$, el intervalo de confianza para la *odds-ratio* al 95 % de nivel de confianza, así como la significación conjunta de todas las covariables del modelo.

Contrastes de significación individual

Por un lado, el estadístico o estadígrafo de Wald permite contrastar la importancia de la aportación de cada variable en el modelo logístico, a través de la hipótesis nula de que su coeficiente β_j sea cero frente a que sea significativamente distinto de cero.

Así, una prueba no significativa indica que la aportación de la variable no es significativamente importante; sin embargo, una prueba significativa reafirmaría la relevancia de la variable en el modelo de regresión logístico.

Observar que se trata de tests de significación individual de cada covariable, y por tanto, si se utiliza como método para seleccionar que covariables son más significativas o relevantes para mantenerse en el modelo de regresión logística hay que tener en cuenta que la reducción y/o inclusión de una covariable en el modelo puede afectar a la significación de las restantes.

Test de significación conjunta

Por otro lado, el contraste de significación conjunta de las covariables consideradas en el modelo de regresión logístico (similar al contraste de regresión en el análisis del modelo de regresión lineal), se establece mediante la hipótesis nula de que todos los coeficientes de las covariables del modelo sean cero, lo que significa que las covariables, en su conjunto, no son relevantes en el modelo. En este caso, el test viene dado a través del estadístico G del logaritmo de razón de verosimilitudes.

Este estadístico $G = D_0 - D$ es la diferencia entre la deviance del modelo sin ninguna covariable y la deviance del modelo completo de regresión logístico, y su distribución aproximada es una chi-cuadrado de Pearson con k grados de libertad (k número de covariables y factores en el modelo).

Selección de covariables

El desarrollo de los procesos de selección de covariables (variables o factores de riesgo) para construir un modelo adecuado de regresión logística con las componentes más relevantes, se basan habitualmente en estos contrastes de significación individual o conjunta, de forma análoga a la selección de un modelo en el análisis de regresión lineal.

Dentro de estos procesos de selección, señalar que las variables ficticias correspondientes a una variable categórica tienen que ser incluidas o eliminadas del modelo en bloque, dado que representan las categorías o valores de un mismo factor. No obstante, también pueden ser redefinida un factor agrupando categorías no significativamente diferentes en cuanto a su influencia en el modelo logístico; en tal caso, este factor modificado sigue formando una partición de la población y también debe ser incluido o eliminado en bloque en un proceso de selección.

Medidas de bondad de ajuste del modelo

En relación a la bondad del modelo de regresión logística, se disponen de diferentes criterios para cuantificar su bondad, así como para contrastar la significación del ajuste del modelo logit.

Como hemos visto, la **deviance** es una medida que acumula las desviaciones en las observaciones producidas por el modelo, $D = \sum_{i=1}^{n} d_i^2$ (desajustes o errores), es decir, representa una medida similar a la suma de cuadrados de los residuos del modelo de regresión lineal. Por ello, podría utilizarse, como medida global de bondad del modelo logístico, la transformación dada por:

$$R^2 = 1 - \frac{D}{D_0}$$

aunque no se trata de un coeficiente de determinación, se mantiene la notación por analogía al coeficiente de determinación de la regresión lineal, dado que en caso de ajuste perfecto $D = 0$ y por tanto $R^2 = 1$, y cuando las covariables no aportan nada $D = D_0$ y por tanto $R^2 = 0$.

Test chi-cuadrado de Pearson

Uno de los contrastes de ajuste más conocido es el chi-cuadrado de Pearson, el cual está basado en la comparación entre los valores observados en la muestra y los pronosticados o esperados en caso de ser cierta la hipótesis nula de buen ajuste.

En nuestro caso, para calcular el estadístico chi-cuadrado de Pearson se realiza la comparación en la partición formada por los grupos de observaciones para los que se obtiene la misma predicción de la probabilidad $\widehat{p}_{\mathbf{x}}$ a través del modelo estimado de regresión logístico, es decir, los grupos vienen determinados por las distintas características $\mathbf{x} = (x_1, ..., x_k)$ que forman el modelo logit.

Así, el estadístico chi-cuadrado de Pearson está determinado por la suma de cuadrados de las frecuencias observadas tipificadas o estandarizadas en esta partición bajo los valores pronosticados, cada término de esta suma se llama residuo de Pearson en ese grupo de observaciones, teniendo una

distribución aproximada chi-cuadrado de Pearson con k^* grados de libertad, con k^* número de grupos menos parámetros en el modelo (constante más covariables).

Test de desviaciones residuales

Una alternativa al contraste de ajuste de Pearson consiste en utilizar las desviaciones residuales o pseudo-residuos $d^2_{\mathbf{x}}$ en vez de los residuos de Pearson, esto es, las diferencias entre la desviación en cada observación del grupo determinado por las covariables $\mathbf{x} = (x_1, ..., x_k)$, a través de la probabilidad estimada $\widehat{p}_{\mathbf{x}}$ y la observada $p_{\mathbf{x}} = \sum_{i|\mathbf{x}} y_i/n_{\mathbf{x}}$ en cada grupo $\mathbf{x}$.

Observar que para cada grupo $\mathbf{x}$, el signo de la desviación residual $d_{\mathbf{x}}$ está determinado por el signo de la diferencia $p_{\mathbf{x}} - \widehat{p}_{\mathbf{x}}$.

Test de Hosmer-Lemeshow

No obstante, estos contrastes pueden presentar inconvenientes derivados del gran número de grupos formados por las distintas características $\mathbf{x} = (x_1, ..., x_k)$; en tal caso, el test de Hosmer-Lemeshow contrasta la bondad de ajuste del modelo estimado de regresión logístico de forma similar a la aplicación del contraste de ajuste chi-cuadrado de Pearson a través de una agrupación de los valores pronosticados de la probabilidad de ocurrencia del suceso de interés. En concreto, el estadístico de Hosmer-Lemeshow se basa en g grupos ordenados de las estimaciones $\widehat{p}_i$ (usualmente $g = 10$ formando una partición de $[0, 1]$), calcula las frecuencias esperadas de cada grupo a través del promedio de las n_k estimaciones que lo componen, $\overline{p}_k$, siendo los n_k equilibrados (iguales o próximos entre sí), y las compara con las frecuencias observadas o_k para los registros correspondientes a cada grupo, utilizando una distribución aproximada chi-cuadrado de Pearson con $g - 2$ grados de libertad.

Los residuos utilizados para los contrastes de ajuste, residuos de Pearson y desviaciones residuales, miden los desajustes entre cada valor pronosticado por el modelo para la probabilidad de ocurrencia del suceso y su valor observado. En este sentido, estos valores y sus gráficos frente a las probabilidades estimadas mediante el modelo logit, pueden analizarse para identificar posibles observaciones anómalas y/o influyentes en el modelo de regresión logístico utilizando los *leverages* como en el caso de la regresión lineal múltiple.

3. Análisis de curvas ROC

3.1. Clasificación y medidas de predicción

Dentro de la finalidad del análisis de regresión logística, se encuentra la clasificación de los individuos según su pertenencia o no al grupo de la población con presencia del suceso de interés. Por ello, una vez establecido el modelo estimado de regresión logístico, interesa saber su capacidad predictiva, es decir, su eficacia para clasificar a nuevos individuos en uno de ambos grupos mediante el valor estimado de su probabilidad de ocurrencia del suceso de interés clínico.

Medidas de concordancia

Las medidas de concordancia indican lo bien que el modelo logit estimado predice los pares de las observaciones disponibles de ambos grupos, a mayor concordancia mejor predice el modelo de regresión logístico.

Una pareja de observaciones de cada grupo se dice concordante si la probabilidad pronosticada para el que presenta el suceso es mayor que para el que no lo presenta. En términos de riesgo relativo, significa que la observación del grupo $Y = 1$ tiene más riesgo relativo de ser clasificada como presencia del suceso frente a la del grupo $Y = 0$, o en términos de *odds-ratio* que es más probable que se clasifique

correctamente a la observación con $Y = 1$ respecto a que se clasifique incorrectamente a la observación con $Y = 0$. Además, la asociación entre estas concordancias y las observaciones indica el grado de capacidad predictiva del modelo.

Matriz de errores de clasificación

La aplicación en el proceso de discriminación del modelo de regresión logístico, es decir, en el proceso de clasificación de los individuos observados, consiste en realizar esta asignación de los individuos a los grupos a través de sus probabilidades pronosticadas de ocurrencia del suceso de interés clínico.

En particular, uno de los procedimientos más habituales consiste en fijar un punto de corte para la probabilidad estimada, por ejemplo $p = 0.5$, a partir del cual si $\widehat{p}_i > 0.5$ se predice que pertenece al grupo de interés, es decir, se estima $\widehat{y}_i = 1$, siendo $\widehat{y}_i = 0$ en caso contrario. Este procedimiento permite construir una tabla de frecuencias llamada matriz de confusión (ver Tabla 1) compuesta por las predicciones correctas y por las predicciones erróneas a través del método clasificación dado por el modelo logit estimado para las observaciones disponibles.

Tabla 1: Matriz de confusión (tabla de frecuencias).

		Clasificación pronosticada		
		Positivo ($\widehat{Y} = 1$)	Negativo ($\widehat{Y} = 0$)	Total observados
Realidad observada	Presencia ($Y = 1$)	VP=Verdaderos Positivos	FN=Falsos Negativos	TP=Total con Presencia
	Ausencia ($Y = 0$)	FP=Falsos Positivos	VN=Verdaderos Negativos	TA=Total con Ausencia
	Total pronosticados	TRP=Total resultado positivo	TRN=Total resultado negativo	n=tamaño muestral

Observar que en este caso, la Tabla 1 recoge las coincidencias y las discrepancias entre la clasificación real registrada para cada observación y la que se deriva de las probabilidades pronosticadas mediante el modelo de regresión logístico. Así, la eficiencia del modelo de regresión logístico para clasificar correctamente a los individuos en el grupo de presencia o en el de ausencia del suceso de interés, es decir, su capacidad para predecir con precisión la variable respuesta dados los valores de los términos (predictores) en el modelo, viene resumida en la Tabla 1, proporcionando una serie de medidas adicionales sobre esta capacidad predictiva.

Sensibilidad y especificidad

Entre estas medidas destaca el par formado por la *sensibilidad* y la *especificidad*:

- La sensibilidad S es la probabilidad de predecir correctamente la pertenencia al grupo del suceso, es decir, $(\widehat{Y} = 1|Y = 1)$, siendo su valor estimado la tasa o fracción de verdaderos positivos
- La especificidad E es la probabilidad de predecir correctamente la no pertenencia al grupo del suceso, es decir, $(\widehat{Y} = 0|Y = 0)$, siendo su valor estimado la tasa o fracción de verdaderos negativos

y este par de medidas de acierto en la clasificación determinan los errores de la misma, dado que la tasa de falsos negativos $(\widehat{Y} = 0|Y = 1)$ es $1 - S$ y la tasa de falsos positivos $(\widehat{Y} = 1|Y = 0)$ es $1 - E$.

Otras medidas de eficiencia del modelo logit estimado vienen dadas por los *valores predictivos (positivo y negativo)* que representan las probabilidades de $(Y = 1|\widehat{Y} = 1)$ e $(Y = 0|\widehat{Y} = 0)$, respectivamente. El *índice de Youden* que representa la diferencia entre las probabilidades de clasificar en el

grupo de presencia ($\widehat{Y} = 1$) correcta e incorrectamente, $\gamma = S - (1 - E)$. Y las *odds* asociadas a la clasificación resultante a través del modelo logístico, dadas por:

- La *odds* de predecir correctamente la presencia, $odds_{VP} = \frac{S}{1-S}$, que indica cuanto más probable es que clasifique correctamente respecto de que clasifique incorrectamente a un individuo que pertenece al grupo de interés ($Y = 1$)
- La *odds* de predecir incorrectamente la ausencia, $odds_{FP} = \frac{1-E}{E}$, que mide cuanto más probable es que clasifique incorrectamente respecto de que clasifique correctamente a un individuo que no pertenece al grupo de interés ($Y = 0$)

y como consecuencia de estas dos últimas, el $odds\text{-}ratio_{(VP,FP)} = \frac{odds_{VP}}{odds_{FP}}$ que representa cuanto más probable es que el modelo logístico clasifique correctamente respecto a que clasifique incorrectamente la pertenencia al grupo de interés.

3.2. Curvas ROC

Observar que las medidas de la capacidad de predecir con precisión del modelo estimado de regresión logístico, son relativas a la clasificación particular realizada a través de este modelo logit, ya que si se modificara el punto de corte $p = 0.5$ se obtendría otra clasificación diferente, y por tanto, una Tabla 1 distinta. Así, se plantea el problema de fijar el punto de corte adecuado para p; por ejemplo, si fuese conocida la proporción real de individuos en la población bajo estudio que pertenecen al grupo de interés, podría variarse p hasta que la proporción del total de resultados positivos TRP sea lo más próxima a la proporción poblacional. Sin embargo, esta proporción poblacional suele ser desconocida en la práctica, siendo habitual recurrir a la construcción de una *curva ROC* para buscar el punto de corte p más apropiado.

La curva ROC se construye representando los pares de puntos $(1-E, S)$ para todos los posibles puntos de corte $p \in [0, 1]$, es decir, la representación de la tasa de verdaderos positivos frente a la tasa de falsos positivos.

Observar que esta gráfica se encuentra en el cuadrado unidad, y algunas veces se presenta en una escala porcentual. Así, el punto $(0, 0)$ de la curva ROC corresponde a una predicción del modelo logit con todos los individuos clasificados como ausencia del suceso (no hay falsos positivos pero tampoco verdaderos positivos); el punto $(1, 1)$ de la curva ROC corresponde a una predicción con todos los individuos clasificados como presencia del suceso (no hay falsos negativos pero tampoco verdaderos negativos); y el punto $(0, 1)$ corresponde a la predicción perfecta con todos los individuos clasificados correctamente (no hay errores: ni falsos positivos, ni falsos negativos).

A partir de esta idea de predicción perfecta, un modelo de regresión logística tendrá la capacidad de clasificar correctamente a todos los individuos de la población si su curva ROC contiene este punto $(0, 1)$, y por consiguiente, el punto de corte óptimo p será el correspondiente a dicho valor de especificidad y sensibilidad. Obviamente, la curva ROC que incluye este punto es la recta horizontal entre $(0, 1)$ y $(1, 1)$, siendo el área bajo esta curva igual a 1. No obstante, en la práctica, una curva ROC no incluye este punto, por lo que el área bajo la curva ROC será inferior a 1 ($AUC < 1$).

Área bajo la curva ROC

En este contexto, la curva ROC asociada a un modelo de regresión logístico representa todas las clasificaciones posibles, a través de sus valores pronosticados, proporcionando una medida global de la capacidad global de discriminación del modelo logit estimado, dada por su AUC, frente a las anteriores medidas asociadas a una clasificación particular.

En general, se asume que $AUC \geq 0.5$, dado que la curva ROC representada por la diagonal principal del cuadrado unidad se corresponde con un $odds\text{-}ratio_{(VP,FP)} = 1$ para todo punto de corte,

es decir, es igualmente probable que una predicción del modelo logístico como pertenencia al grupo de interés sea correcta o incorrecta, lo que se interpreta como una clasificación al azar o aleatoria, siendo $AUC = 0.5$ el área de esta curva ROC. Por otro lado, si fuese $AUC < 0.5$ indicaría que en la mayor parte de la curva se tendría $odds\text{-}ratio_{(VP,FP)} < 1$, es decir, sería menos probable que la clasificación en el grupo del suceso fuese correcta que incorrecta, y en tal caso, se mejoraría intercambiando el sentido de la clasificación realizada a través del modelo logístico estimado. Además, un valor de AUC cercano a 1 mostrará su proximidad a la discriminación perfecta (sin errores).

Un mecanismo para valorar e interpretar la proximidad del AUC a 1 con respecto a la capacidad de predecir con precisión del modelo estimado de regresión logístico, corresponde al criterio general de Hosmer y Lemeshow (2000) dado en la Tabla 2.

Tabla 2: Categorías del AUC para la capacidad de discriminar.

$AUC = 0.5$	No discrimina (como lanzar una moneda)
$0.7 \leq AUC < 0.8$	Discriminación aceptable
$0.8 \leq AUC < 0.9$	Discriminación excelente
$0.9 \leq AUC$	Discriminación extraordinaria

Punto de corte óptimo

Un procedimiento para elegir un punto de corte óptimo, para la clasificación o discriminación a través del modelo estimado de regresión logístico, consiste en representar la curva ROC correspondiente a las clasificaciones con todos los posibles puntos de corte, seleccionando aquél que en algún sentido más se aproxime a la clasificación perfecta, es decir, cuyos valores de especificidad y sensibilidad correspondan al punto $(1 - E, S)$ más próximo a $(0, 1)$. Por ejemplo, gráficamente puede trazarse una recta paralela a la diagonal principal (esto es, de pendiente 1) seleccionando aquella que sea tangente a la curva y más próxima a $(0, 1)$.

Entre las diferentes técnicas utilizadas para fijar un punto de corte óptimo, una de las más habituales y sencillas consiste en seleccionar el punto de intersección entre la curva ROC y la diagonal opuesta del cuadrado unidad, ya que en dicha intersección son iguales la especificidad y la sensibilidad, $S = E$, y por tanto, en este punto se mantiene una relación de compensación adecuada entre las correctas e incorrectas predicciones de la presencia del suceso de interés (la mismas proporciones de predicciones correctas tanto positivas como negativas, y la misma proporción de predicciones incorrectas). Además, este punto puede obtenerse rápidamente a través de la gráfica conjunta de la especificidad (verdaderos negativos) y de la sensibilidad (verdaderos positivos) frente a los posibles puntos de corte p, puesto que la primera curva es creciente y la segunda decreciente respecto del punto de corte.

4. Caso práctico

Veamos un caso práctico de aplicación del análisis de regresión logística a través del programa R. En este estudio se analiza el bajo peso de los recién nacidos bajo determinados factores de riesgo, el conjunto de datos observados se encuentra en el texto de Hosmer y Lemeshow (2000) y fueron recogidos en el Baystate Medical Center de Massachusetts, durante 1986 para intentar identificar los factores que contribuyen a aumentar el riesgo de bebés nacidos con bajo peso. Se recogieron 189 observaciones de mujeres de las cuales 59 tuvieron bebés con bajo peso al nacer.

Las variables (columnas) del fichero de datos birthwt, se describen la Tabla 3, y los datos se encuentran en la librería *MASS* de R.

En primer lugar, consideraremos un modelo de regresión logístico simple (una sola covariable) para mostrar los resultados del análisis; posteriormente, ampliaremos el modelo con un factor de riesgo

Tabla 3: Descripción de las variables del archivo birthwt.

Nombre	Descripción
low	Indicador de bajo peso al nacer (1=bajo peso<2.5kg, 0=normal)
age	Edad de la madre en años en el momento del parto
lwt	Peso de la madre (en *pounds*=0.45359237Kg) en el último periodo menstrual
race	Raza de la madre (1=blanca, 2=negra, 3=otra)
smoke	Indicador de madre fumadora en el embarazo (1=fumadora, 0=no fumadora)
ptl	Número de partos prematuros anteriores
ht	Indicador de hipertensión (1=si, 0=no)
ui	Indicador de presencia de irritabilidad uterina (1=si, 0=no)
ftv	Número de visitas al centro médico durante el primer trimestre
bwt	Peso del bebé al nacer en gramos

cualitativo (factor categórico) destacando su tratamiento en el análisis de regresión logística, y por último, analizaremos el modelo más completo seleccionando los factores que se consideren relevantes para la regresión logística.

Ejemplo de regresión logística simple

En este ejemplo analizaremos el modelo de regresión logístico para la variable respuesta *low* que determina el suceso de interés clínico ($low = 1$ recién nacido con bajo peso) a través de la covariable *lwt* (peso de la madre en el último periodo menstrual). Para ello, abrimos el fichero de datos y declaramos los factores, observar que en la declaración de un factor puede indicarse una etiqueta para sus niveles que facilitará la posterior interpretación de los resultados, como ejemplo se muestran las primeras filas de la salida de la codificación:

```
> data(birthwt)
> attach(birthwt)
> race <- factor(race,labels=c("blanca","negra","otra"))
> ht <- factor(ht,labels=c("no","si"))
> smoke <- factor(smoke,labels=c("no","si"))
> ui <- factor(ui,labels=c("no","si"))
> datos <- data.frame(low,age,lwt,race,smoke,ptl,ht,ui,ftv)
> datos
    low age lwt   race smoke ptl ht ui ftv
1     0  19 182  negra    no   0 no si   0
2     0  33 155   otra    no   0 no no   3
3     0  20 105 blanca    si   0 no no   1
4     0  21 108 blanca    si   0 no si   2
```

Ahora, utilizando la función *glm* de modelo lineal general de R, como se muestra debajo, se obtiene el modelo estimado de regresión logística, el cual se muestra llamando al modelo asignado o más completo con la función *summary*:

```
> modelo1 <- glm(low~lwt,family=binomial); modelo1

Call:  glm(formula = low ~ lwt, family = binomial)

Coefficients:
(Intercept)          lwt
    0.99831     -0.01406
```

```
Degrees of Freedom: 188 Total (i.e. Null);  187 Residual
Null Deviance:     234.7
Residual Deviance: 228.7  AIC: 232.7
> summary(modelo1)

Call:
glm(formula = low ~ lwt, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0951  -0.9022  -0.8018   1.3609   1.9821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99831    0.78529   1.271   0.2036
lwt         -0.01406    0.00617  -2.279   0.0227 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 228.69  on 187  degrees of freedom
AIC: 232.69

Number of Fisher Scoring iterations: 4
```

Entre estos resultados del análisis de regresión logística para bebé de bajo peso *low* frente al peso de la madre *lwt*, destacamos que el modelo logístico estimado es:

$$\log \widehat{odds}_{lwt} = \log\left(\frac{\widehat{p}_{lwt}}{1-\widehat{p}_{lwt}}\right) = 0.99831 - 0.01406 \cdot lwt \;\Leftrightarrow\; \widehat{p}_{lwt} = \frac{e^{0.99831-0.01406\cdot lwt}}{1+e^{0.99831-0.01406\cdot lwt}}$$

y los intervalos de confianza de los coeficientes o de los *odds-ratio*:

```
> confint(modelo1)
                  2.5 %       97.5 %
(Intercept) -0.48116701  2.611748138
lwt         -0.02696198 -0.002650036
> exp(modelo1$coefficients)
(Intercept)         lwt
  2.7137035   0.9860401
> exp(confint(modelo1))
                2.5 %     97.5 %
(Intercept) 0.6180617 13.6228447
lwt         0.9733982  0.9973535
```

Además, la estimación del coeficiente el peso de la madre en la última menstruación, $\widehat{\beta}_{lwt} = -0.01405826$, cuyo $\widehat{odds\text{-}ratio}_{lwt} = e^{-0.01405826} = 0.9860401$, y el P-valor 0.0227 muestra que el coeficiente β_{lwt} es significativamente distinto de cero, es decir, aunque $\widehat{odds\text{-}ratio}_{lwt} = 0.986$ esté próximo a 1 si resulta ser significativamente distinto de 1. Esto indicaría que si influye la variable *lwt* pues al aumentar en una libra el peso de la madre en la última menstruación se reduce ligeramente la probabilidad de bajo peso del recién nacido frente a un peso normal. Sin embargo, el intercepto o constante resulta no significativo.

En el resumen de los resultados del ajuste del modelo logístico, también se incluyen los términos de deviance con los que se realiza el contraste de significación conjunto, el término *Null Deviance* corresponde al modelo nulo o constante (sin covariables), y el término *Residual Deviance* corresponde al modelo ajustado o estimado. No obstante, podemos indicar en R que realice automáticamente este contraste mediante la comparación de ambos modelos con la función *anova*, e incluso eliminando el efecto del intercepto sobre las covariables con la función *drop1*, en ambos casos eligiendo el tipo de test en el argumento:

```
> anova(modelo1,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: low

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                  188     234.67
lwt   1   5.9813       187     228.69  0.01446 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> drop1(modelo1,test="Chisq")
Single term deletions

Model:
low ~ lwt
       Df Deviance    AIC    LRT Pr(>Chi)
<none>      228.69 232.69
lwt     1   234.67 236.67 5.9813  0.01446 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observamos que en este caso, al incluir sólo una covariable, el intercepto sólo incluye efectos de dicha covariable, por lo que no afecta a la significación conjunta de las covariables; sin embargo, si afectará en los problemas de más de una covariable o factor. Así, esta significación de *lwt* en el modelo, también viene establecida por el P-valor 0.01446 del contraste de significación del conjunto de variables del modelo dado por el estadístico G del logaritmo de razón de verosimilitudes.

Otra forma de obtener el modelo de regresión logística es a través de la librería *rms* de R. Si ejecutamos las funciones que se incluyen debajo, además de algunos de los resultados obtenidos antes, muestra diversas medidas de asociación relacionadas con las concordancias del modelo logístico en la predicción del suceso de interés (bajo peso al nacer), entre otras, la tau de Kendall y la gamma de Goodman y Kruskal:

```
> library("rms")
> datosdist <- datadist(datos)
> options(datadist="datosdist")
> modelo1b <- lrm(low~lwt,y=T,x=T); modelo1b

Logistic Regression Model

lrm(formula = low ~ lwt, x = T, y = T)
```

```
                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes          Indexes
Obs           189    LR chi2      5.98    R2       0.044    C       0.613
 0            130    d.f.            1    g        0.452    Dxy     0.226
 1             59    Pr(> chi2) 0.0145    gr       1.571    gamma   0.232
max |deriv| 5e-08                         gp       0.088    tau-a   0.098
                                          Brier    0.208

          Coef    S.E.   Wald Z Pr(>|Z|)
Intercept  0.9983 0.7853  1.27  0.2036
lwt       -0.0141 0.0062 -2.28  0.0227

> summary(modelo1b)
             Effects              Response : low

 Factor      Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 lwt         110 140  30    -0.42  0.19 -0.78      -0.06
  Odds Ratio 110 140  30     0.66    NA  0.46       0.94
> anova(modelo1b)
                Wald Statistics          Response: low

 Factor     Chi-Square d.f. P
 lwt        5.19       1    0.0227
 TOTAL      5.19       1    0.0227
```

otra ventaja de ajustar un modelo de regresión logística a través de la función *lrm*, consiste en la posibilidad un contraste de bondad de ajuste de forma automática, aunque podrían realizarse a partir de los residuos en los que se basan estos contrastes, por ejemplo añadiendo en la tabla de datos los residuos de Pearson o los residuos en la predicción de la respuesta, así como las probabilidades estimadas de ocurrencia con el modelo logístico ajustado:

```
> datos$res.pearson <- residuals(modelo1,type="pearson")
> datos$res.respuesta <- residuals(modelo1,type="response")
> datos$prob.estimadas <- modelo1$fitted.values
```

la función *lrm* permite indicar el argumento "*gof*" (goodness of fit) para realizar un contraste de bondad de ajuste:

```
> residuals(modelo1b,"gof")
Sum of squared errors     Expected value|H0                    SD
           39.3018057             39.3859060             0.1223962
                    Z                       P
           -0.6871146              0.4920105
```

Notar que en este caso, el diagnóstico del modelo se reduce a las medidas y contrastes de bondad de ajuste, pues no se imponen condiciones iniciales de homoscedasticidad o normalidad.

En cualquier caso, para mostrar como aplicar el modelo estimado de regresión logística en la clasificación y medir su capacidad predictiva, considerando el punto de corte $p = 0.5$ a partir del cual se predice la pertenencia al grupo de interés (bebé pronosticado de bajo peso), se representan las frecuencias de predicciones correctas e incorrectas de pertenencia a cada grupo en una tabla de doble entrada (matriz de confusión). A partir del modelo estimado y los resultados obtenidos en R, podemos elaborar esta tabla con la función *clasificacion* incluida en el fichero RData como sigue:

```
> clasificacion(y=low, yhat=modelo1, corte=0.5)
```

	Clasifica correcto	Clasifica incorrecto	Clasifica total
1	0	0	0
0	130	59	189
Total	130	59	189

donde se observa que este punto de corte no es adecuado para la clasificación aunque no pronostica erróneamente casos de bebés de bajo peso, dado que predice peso normal para todos los recién nacidos.

En este contexto, vemos la necesidad de aplicar algún mecanismo que permita medir la capacidad de clasificación global del modelo de regresión logística ajustado y seleccionar un punto de corte más apropiado. Para ello, aplicaremos el Análisis de Curvas ROC a los valores pronosticados por el modelo logístico (probabilidades estimadas de ocurrencia del suceso de interés). Por ejemplo, utilizando el paquete *ROCR* de R, se representa fácilmente la curva ROC correspondiente a todos las posibles clasificaciones en bebé de bajo peso o normal según los distintos puntos de corte (Figura 1, y entre los resultados ofrece la medida del área bajo la curva AUC sobre la capacidad global de clasificación del modelo logístico estimado. Como se ve a continuación, en este caso la capacidad global de precisión en la discriminación dada por el área bajo la curva ROC, $AUC = 0.613$, se queda por debajo de la capacidad aceptable por el criterio de Hosmer-Lemeshow de la Tabla 2.

```
> library("ROCR")
> predicciones <- prediction(datos$prob.estimadas,datos$low)
> curva <- performance(predicciones,"tpr","fpr")
> plot(curva,colorize=TRUE,main="Curva ROC")
> performance(predicciones,"auc")
Slot "y.name":
[1] "Area under the ROC curve"

Slot "y.values":
[[1]]
[1] 0.613103
```

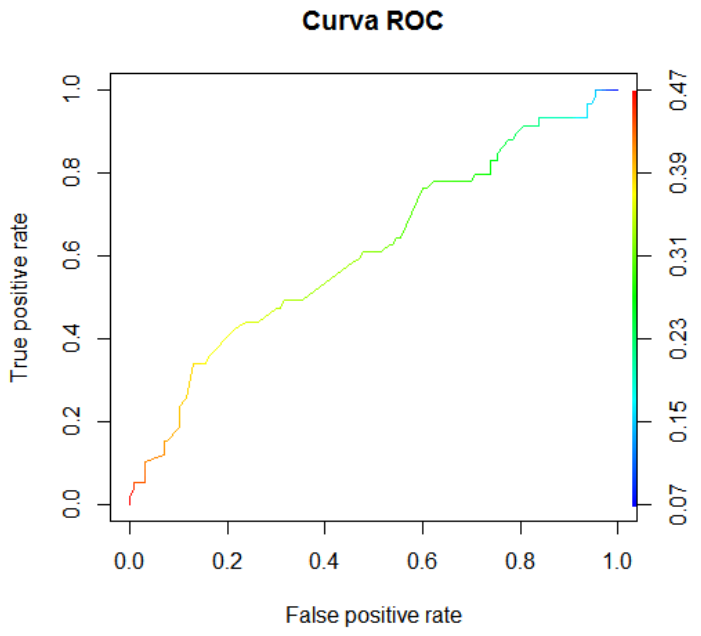

Figura 1: Gráfica de la curva ROC del modelo de regresión logística

Para finalizar, utilizaremos la librería *Epi* de R, que nos proporciona un mecanismo gráfico de obtención de un punto de corte óptimo para la clasificación mediante el modelo de regresión logística estimado. Así, con la función *ROC* de este paquete y los argumentos definidos como sigue, los resultados más relevantes del análisis de la curva se incluyen en la propia gráfica que realiza en la Figura 2:

```
> library("Epi")
> curva2 <- ROC(form=low~lwt,plot="ROC",PV=TRUE,MX=TRUE,AUC=TRUE,data=datos)
```

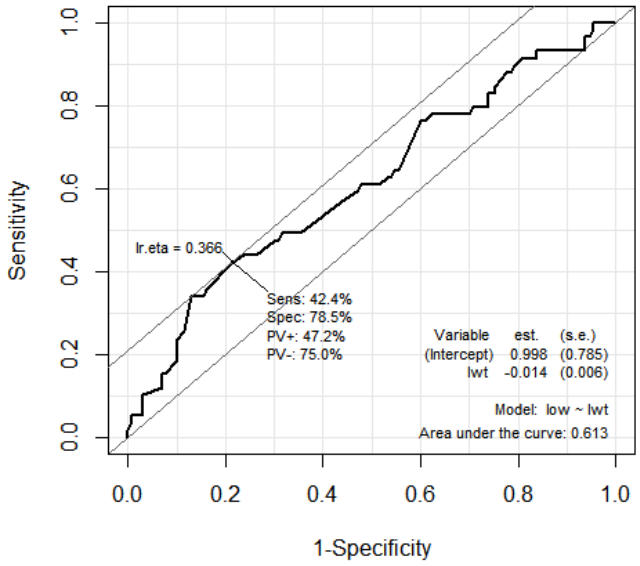

Figura 2: Gráfica de la curva ROC y punto de corte

entre estos resultados, el área bajo la curva ROC dada anteriormente, y el punto de corte seleccionado $p = 0.366$ junto con los valores de sensibilidad y especificidad correspondientes a la clasificación a partir de los valores estimados de la probabilidad de bebé con bajo peso. También puede obtenerse con:

```
> curva2$AUC
[1] 0.613103
```

Además, puede crearse un nuevo cuadro de datos con los principales valores de la curva ROC, sensibilidad y especificidad, junto con sus correspondientes puntos de corte, de modo que en la propia tabla de datos de RStudio puede comprobarse el cruce entre estos dos valores, dando otro procedimiento de elección de un punto de corte óptimo (entre 0.331 y 0.334):

```
> modelo1sens <- curva2$res$sens; modelo1espec <- curva2$res$spec
> modelo1pcortes <- curva2$res$lr.eta
> modelo1ROC <- data.frame(modelo1pcortes,modelo1sens,modelo1espec)
> viewData(modelo1ROC)

      modelo1pcortes   modelo1sens   modelo1espec
48       0.33120606    0.57627119    0.553846154
49       0.33432744    0.49152542    0.646153846
```

este cruce entre los valores de la sensibilidad y especificidad se aprecia en la siguiente Figura 3 realizada cambiando los argumentos de la función anterior, aunque en este caso se visualiza mejor utilizando las funciones gráficas de R sobre nuestro marco de datos:

```
> ROC(form=low~lwt,plot="sp")
> plot(modelo1pcortes,modelo1sens,type="l",xlab="Puntos de corte",
 + ylab="                                          S",
+ main="Curvas de Sensibilidad y Especificidad")
> par(new=TRUE)
> plot(modelo1pcortes,modelo1espec,type="l",xlab="",
 + ylab="E                                        ")
```

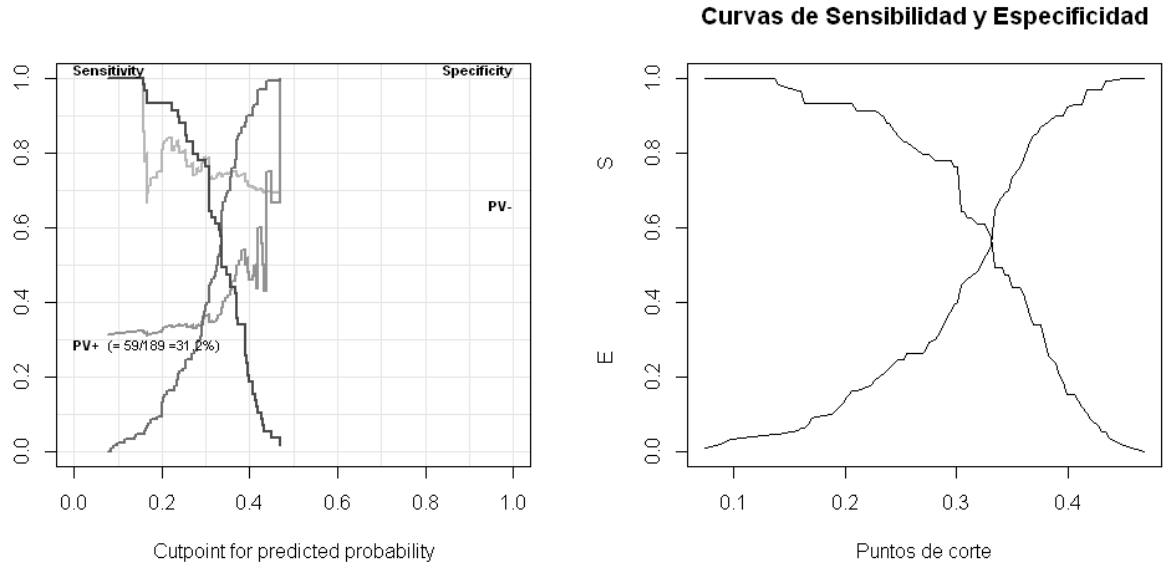

Figura 3: Gráficas de sensibilidad y especificidad respecto de los puntos de corte

Y ahora, repetimos la tabla o matriz de confusión para las clasificaciones correctas e incorrectas realizadas a partir del modelo de regresión logística mediante el punto de corte óptimo $p = 0.366$:

```
> clasificacion(y=low, yhat=modelo1, corte=0.366)
      Clasifica correcto Clasifica incorrecto Clasifica total
1                     25                   28              53
0                    102                   34             136
Total                127                   62             189
```

cuya sensibilidad y especificidad se indica en la Figura 2.

Ejemplo de regresión logística con un factor categórico

Veamos ahora como realizar el análisis de regresión logística considerando más de una variable en el modelo, siendo alguna de ellas categórica (variable cualitativa). Para ello, consideraremos la inclusión del factor *race* (formada por tres clases: 1=blanca, 2=negra, 3=otra) en el modelo logit anterior para analizar su importancia sobre la variable respuesta *low* y para mejorar la predicción del suceso $low = 1$, bajo peso del recién nacido, obtenida en el caso simple a través de la covariable *lwt*. Y para que resulte más cómodo, utilizaremos la misma notación que en el ejemplo anterior.

```
> modelo1 <- glm(low~lwt+race,family=binomial); modelo1

Call:  glm(formula = low ~ lwt + race, family = binomial)

Coefficients:
```

```
(Intercept)          lwt    racenegra     raceotra
    0.80575     -0.01522      1.08107      0.48060

Degrees of Freedom: 188 Total (i.e. Null);  185 Residual
Null Deviance:     234.7
Residual Deviance: 223.3  AIC: 231.3
> summary(modelo1)

Call:
glm(formula = low ~ lwt + race, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3491  -0.8919  -0.7196   1.2526   2.0993

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.805753   0.845167   0.953   0.3404
lwt         -0.015223   0.006439  -2.364   0.0181 *
racenegra    1.081066   0.488052   2.215   0.0268 *
raceotra     0.480603   0.356674   1.347   0.1778
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 223.26  on 185  degrees of freedom
AIC: 231.26

Number of Fisher Scoring iterations: 4
```

y añadimos el cálculo de los intervalos de confianza para los coeficientes y los *odds-ratio*:

```
> confint(modelo1)
                 2.5 %       97.5 %
(Intercept) -0.79971278  2.527594704
lwt         -0.02862628 -0.003269207
racenegra    0.11991221  2.049646955
raceotra    -0.21917539  1.183795501
> exp(modelo1$coefficients)
(Intercept)         lwt   racenegra    raceotra
  2.2383824   0.9848922   2.9478208   1.6170496
> exp(confint(modelo1))
                2.5 %     97.5 %
(Intercept) 0.4494580 12.5233475
lwt         0.9717796  0.9967361
racenegra   1.1273979  7.7651592
raceotra    0.8031808  3.2667497
```

Aplicar también las funciones *anova* y *drop1* en este modelo logístico y comentar los resultados.

```
> anova(modelo1,test="Chisq")
> drop1(modelo1,test="Chisq")
```

Asimismo, aplicando la función *lrm* del paquete *rms* de R, se obtienen algunas medidas de asociación relacionadas con las concordancias del modelo y la significación conjunta:

```
> modelo1b <- lrm(low~lwt+race,y=T,x=T); modelo1b

Logistic Regression Model

lrm(formula = low ~ lwt + race, x = T, y = T)

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes          Indexes
Obs           189    LR chi2     11.41    R2       0.082    C       0.647
 0            130    d.f.            3    g        0.643    Dxy     0.293
 1             59    Pr(> chi2) 0.0097    gr       1.903    gamma   0.296
max |deriv| 2e-07                         gp       0.128    tau-a   0.127
                                          Brier    0.202

           Coef    S.E.   Wald Z Pr(>|Z|)
Intercept   0.8058 0.8452  0.95  0.3404
lwt        -0.0152 0.0064 -2.36  0.0181
race=negra  1.0811 0.4881  2.22  0.0268
race=otra   0.4806 0.3567  1.35  0.1778

> summary(modelo1b)
             Effects              Response : low

 Factor              Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 lwt                 110 140  30    -0.46  0.19 -0.84      -0.08
  Odds Ratio         110 140  30     0.63    NA  0.43       0.92
 race - negra:blanca   1   2  NA     1.08  0.49  0.12       2.04
  Odds Ratio           1   2  NA     2.95    NA  1.13       7.67
 race - otra:blanca    1   3  NA     0.48  0.36 -0.22       1.18
  Odds Ratio           1   3  NA     1.62    NA  0.80       3.25
> anova(modelo1b)
                Wald Statistics          Response: low

 Factor     Chi-Square d.f. P
 lwt         5.59      1    0.0181
 race        5.40      2    0.0671
 TOTAL      10.13      3    0.0175
```

Los resultados de este análisis se presentan de la misma forma que en el caso anterior, observando que en la significación individual y conjunta de los términos del modelo logístico, al incluir el factor *race*, se desglosan sus coeficientes en dos, presentando la significación y las *odds-ratio* para cada uno de ellos. Como hemos mencionado, no sería conveniente incluir este factor como covariable, dado que sus valores numéricos eran identificativos de sus clases pero no cuantificadores de las mismas, lo que provocaría una incorrecta interpretación de su aportación en el modelo. En este caso, la decisión de incluir o no este factor en el modelo debe considerarse en bloque con todos sus clases, siendo 0.0268 el P-valor de la clase raza negra, lo que indica que es significativo en el modelo, y 0.1778 para la categoría otras razas. Aunque este último podría considerarse que no interviene de forma significativa en el modelo, debe tenerse en cuenta que la $\widehat{odds\text{-}ratio}_{negra,blanca} = 2.9478$ indica que es casi 3 veces más

probable que ocurra que el recién nacido tenga bajo peso frente a que sea normal cuando la madre es de raza negra respecto de que sea de raza blanca; y $\widehat{odds\text{-}ratio}_{otra,blanca} = 1.617$ indica que también es 1.6 veces más probable que sea de bajo peso frente a normal en madres de otras razas con respecto a las madres de raza blanca, aunque esta última no es significativamente distinto de 1 (podría considerarse la agrupación entre estos dos grupos de razas). Además, mantener este factor en el modelo no altera significativamente la importancia de la covariable *lwt*, dado que prácticamente mantiene su *odds-ratio* e incluso se mejora ligeramente su significación individual y la del conjunto de variables del modelo (ambos términos *lwt* y *race*) con un P-valor de 0.0097 en el test de razón de verosimilitudes G, siendo el modelo estimado de regresión logística

$$\widehat{p}_{(lwt,race)} = \frac{e^{0.805753-0.0152231\cdot lwt+1.08107\cdot race_2+0.480603\cdot race_3}}{1+e^{0.805753-0.0152231\cdot lwt+1.08107\cdot race_2+0.480603\cdot race_3}}.$$

En cuanto a los resultados de las pruebas de ajuste de este modelo estimado, además del contraste utilizado en el ejemplo simple, si está disponible la librería *ResourceSelection*, puede utilizarse la función *hosmer.test* del test de ajuste de Hosmer y Lemeshow:

```
> residuals(modelo1b,"gof")
Sum of squared errors     Expected value|H0                    SD
           38.2268160            38.2138614             0.1733477
                    Z                     P
            0.0747321             0.9404279
> library("ResourceSelection")
> hoslem.test(low, fitted(modelo1))

Hosmer and Lemeshow goodness of fit (GOF) test

data:  low, fitted(modelo1)
X-squared = 9.8031, df = 8, p-value = 0.2791
```

observando que los tests indican que puede asumirse un buen ajuste. No obstante, hay que tener en cuenta que cuando aumenta el número de clases en la aplicación de los test de ajuste chi-cuadrado (se aproxima al número de observaciones) pierden eficacia por reducirse el tamaño de cada grupo, siendo en este caso más fiable la aplicación del test de Hosmer-Lemeshow. Asimismo, se observa que han mejorado ligeramente con respecto al modelo anterior con sólo una covariable las medidas de asociación relacionadas con las concordancias.

Por otro lado, para concluir el ejemplo realizaremos la clasificación producida por el modelo logístico ajustado a través de un análisis de su curva ROC y eligiendo un punto de corte apropiado para discriminar entre los bebés que sean más probables de tener bajo peso:

```
curva2 <- ROC(form=low~lwt+race,plot="ROC",PV=TRUE,MX=TRUE,AUC=TRUE,data=datos)
```

donde observamos que ha aumentado ligeramente la capacidad de discriminar correctamente el modelo logístico al incluir el factor raza, $AUC = 0.647$, y si repetimos la tabla o matriz de confusión para las clasificaciones correctas e incorrectas realizadas a partir del modelo de regresión logística mediante el punto de corte óptimo $p = 0.408$, como indica la gráfica de la Figura 4, se tiene:

```
> clasificacion(y=low, yhat=modelo1, corte=0.408)
      Clasifica correcto Clasifica incorrecto Clasifica total
1                     21                   16              37
0                    114                   38             152
Total                135                   54             189
```

para las que se obtienen las medidas de sensibilidad y especificidad también incluidas en la Figura 4.

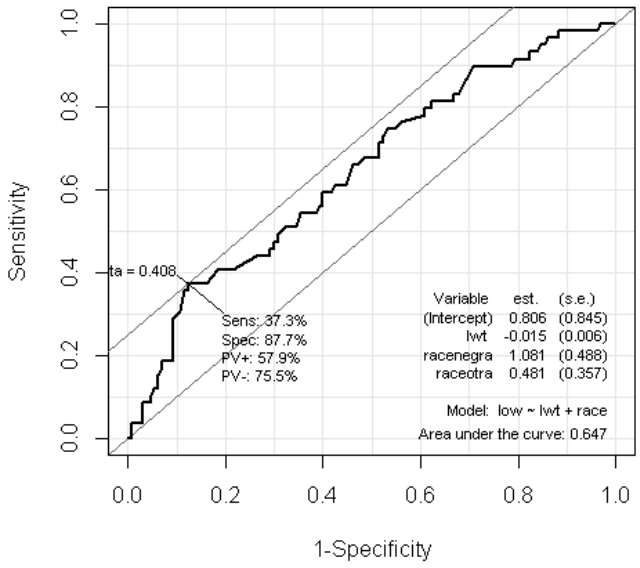

Figura 4: Gráfica de la curva ROC y punto de corte

Ejercicio 4.1 *Crear un cuadro de datos con los valores de sensibilidad y especificidad correspondientes a los posibles puntos de corte para discriminar entre recién nacidos con bajo peso y con peso normal a través de la raza y peso de la madre antes del embarazo.*

Encontrar el punto de corte correspondiente al cruce entre la sensibilidad y especificidad, y realizar una gráfica ilustrativa para visualizar dicho cruce.

Obtener, para dicho punto de corte, la tabla de clasificaciones correctas e incorrectas en la discriminación entre bajo peso y normal con el modelo logístico estimado.

```
> modelo1sens <- curva2$res$sens; modelo1espec <- curva2$res$spec
> modelo1pcortes <- curva2$res$lr.eta
> modelo1ROC <- data.frame(modelo1pcortes,modelo1sens,modelo1espec)
> viewData(modelo1ROC)
> ROC(form=low~lwt+race,plot="sp")
```

Ejemplo de regresión logística múltiple y factores relevantes

En este apartado vemos un ejemplo de selección de los factores más relevantes en el modelo de regresión logístico. Hay diversas técnicas aplicables para desarrollar un proceso de selección del modelo más adecuado, de forma similar a las utilizadas en el modelo de regresión lineal múltiple.

En este sentido, consideramos como modelo logit inicial (constante o nulo) y el modelo logit completo, es decir, compuesto por todas las covariables y factores disponibles en este conjunto de observaciones, incluyendo todos los aspectos de este modelo completo, para analizar la importancia y efecto de una covariable o factor al suprimirla del modelo logit, llamando *modelo0* al primero y manteniendo la notación de *modelo1* para el completo:

```
> modelo0 <- glm(low~1,family=binomial,data=datos); modelo0

Call:  glm(formula = low ~ 1, family = binomial, data = datos)

Coefficients:
```

```
(Intercept)
      -0.79

Degrees of Freedom: 188 Total (i.e. Null);  188 Residual
Null Deviance:     234.7
Residual Deviance: 234.7  AIC: 236.7
> modelo1 <- glm(low~.,family=binomial,data=datos)
> modelo1

Call:  glm(formula = low ~ ., family = binomial, data = datos)

Coefficients:
(Intercept)          age          lwt    racenegra     raceotra      smokesi
    0.48062     -0.02955     -0.01542      1.27226      0.88050      0.93885
        ptl         htsi         uisi          ftv
    0.54334      1.86330      0.76765      0.06530

Degrees of Freedom: 188 Total (i.e. Null);  179 Residual
Null Deviance:     234.7
Residual Deviance: 201.3  AIC: 221.3
```

En primer lugar, analizamos el modelo completo ajustado formado por los dos términos anteriores, *lwt* y *race*, junto con los restantes *age*, *smoke ptl*, *ht*, *ui* y ftv, teniendo en cuenta si alguno de ellos es un factor categórico (como se declararon al crear el cuadro de datos) y mantenemos como covariables los factores cualitativos ordinales. Como se observa en los resultados anteriores, por ejemplo para el factor *smoke* sólo se incluye el término que indentifica a la madre fumadora *smokesi*, dado que su coeficiente representa la importancia de este factor en el modelo al comparar su aportación de ser fumadora respecto de no serlo (caso contrario) en la predicción de bajo peso del recién nacido.

Asimismo, en este análisis de regresión logística, cuyos resultados se incluyen a continuación, se observa que el test de razón de verosimilitudes conjunto del estadístico G es más significativo (p-valor 0.0001). Sin embargo, también se puede comprobar que algunos términos del modelo no son significativos; por ejemplo, la cantidad de visitas al centro médico durante el primer trimestre de embarazo tiene el mayor P-valor en los contrastes individuales, siendo 0.70484, lo que indica que ftv es la covariable menos relevante del modelo (primera candidata a suprimir). También, se observa que el intervalo de confianza al nivel 95 % para su *odds-ratio*$_{ftv}$ incluye el punto 1 entre sus valores centrales, indicando que es admisible que es igualmente probable que el recién nacido sea de bajo peso frente a que su peso sea normal cuando se aumenta en una las visitas al centro médico en el primer trimestre. Comentarios similares pueden realizarse sobre otros términos del modelo.

```
> summary(modelo1)

Call:
glm(formula = low ~ ., family = binomial, data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8946  -0.8212  -0.5316   0.9818   2.2125

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.480623   1.196888   0.402  0.68801
age         -0.029549   0.037031  -0.798  0.42489
lwt         -0.015424   0.006919  -2.229  0.02580 *
```

```
racenegra    1.272260   0.527357   2.413  0.01584 *
raceotra     0.880496   0.440778   1.998  0.04576 *
smokesi      0.938846   0.402147   2.335  0.01957 *
ptl          0.543337   0.345403   1.573  0.11571
htsi         1.863303   0.697533   2.671  0.00756 **
uisi         0.767648   0.459318   1.671  0.09467 .
ftv          0.065302   0.172394   0.379  0.70484
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.28  on 179  degrees of freedom
AIC: 221.28

Number of Fisher Scoring iterations: 4
> confint(modelo1)
                  2.5 %      97.5 %
(Intercept) -1.84121370  2.87745202
age         -0.10373422  0.04207540
lwt         -0.02978452 -0.00246483
racenegra    0.24166064  2.32608774
raceotra     0.02661178  1.76511921
smokesi      0.16158429  1.74790611
ptl         -0.12346116  1.24603059
htsi         0.53239257  3.32119843
uisi        -0.14356295  1.67090307
ftv         -0.28308378  0.39881567
> exp(modelo1$coefficients)
(Intercept)         age         lwt   racenegra    raceotra     smokesi
  1.6170819   0.9708833   0.9846941   3.5689085   2.4120956   2.5570281
        ptl        htsi        uisi         ftv
   1.7217428  6.4449886   2.1546928   1.0674812
> exp(confint(modelo1))
                2.5 %     97.5 %
(Intercept) 0.1586248 17.7689406
age         0.9014649  1.0429731
lwt         0.9706547  0.9975382
racenegra   1.2733620 10.2378101
raceotra    1.0269690  5.8422688
smokesi     1.1753715  5.7425658
ptl         0.8838560  3.4765158
htsi        1.7030020 27.6935195
uisi        0.8662663  5.3169672
ftv         0.7534567  1.4900589

> modelo1b <- lrm(low~.,y=T,x=T,data=datos); modelo1b

Logistic Regression Model

lrm(formula = low ~ ., data = datos, x = T, y = T)
```

```
                    Model Likelihood     Discrimination    Rank Discrim.
                       Ratio Test           Indexes          Indexes
Obs           189    LR chi2     33.39    R2       0.228    C       0.746
 0            130    d.f.            9    g        1.180    Dxy     0.492
 1             59    Pr(> chi2) 0.0001    gr       3.254    gamma   0.493
max |deriv| 2e-04                         gp       0.214    tau-a   0.212
                                          Brier    0.179

           Coef    S.E.   Wald Z Pr(>|Z|)
Intercept   0.4806 1.1969  0.40  0.6880
age        -0.0295 0.0370 -0.80  0.4249
lwt        -0.0154 0.0069 -2.23  0.0258
race=negra  1.2723 0.5274  2.41  0.0158
race=otra   0.8805 0.4408  2.00  0.0458
smoke=si    0.9388 0.4022  2.33  0.0196
ptl         0.5433 0.3454  1.57  0.1157
ht=si       1.8633 0.6975  2.67  0.0076
ui=si       0.7676 0.4593  1.67  0.0947
ftv         0.0653 0.1724  0.38  0.7048

> anova(modelo1b)
                Wald Statistics          Response: low

 Factor     Chi-Square d.f. P
 age         0.64      1    0.4249
 lwt         4.97      1    0.0258
 race        7.12      2    0.0285
 smoke       5.45      1    0.0196
 ptl         2.47      1    0.1157
 ht          7.14      1    0.0076
 ui          2.79      1    0.0947
 ftv         0.14      1    0.7048
 TOTAL      25.70      9    0.0023
```

En relación a las pruebas de ajuste del modelo de regresión logístico completo:

```
> residuals(modelo1b,"gof")
Sum of squared errors     Expected value|H0                    SD
           33.8423832             33.6746219             0.3344046
                    Z                       P
            0.5016718              0.6158984
> hoslem.test(low,fitted(modelo1))

Hosmer and Lemeshow goodness of fit (GOF) test

data:  low, fitted(modelo1)
X-squared = 3.9434, df = 8, p-value = 0.8622
```

se obtienen altas probabilidades de error al rechazar el ajuste, p-valores superiores al 0.05, lo que indica un buen ajuste de las probabilidades pronosticadas por el modelo logístico a las registradas en las observaciones. También, se observa aumento en las medidas de asociación obtenidas anteriormente sobre ambas probabilidades (pronosticadas y observadas).

Además, aplicando la función del análisis de la curva ROC correspondiente a este modelo de regresión logística completo, se obtiene la gráfica de la Figura 5, la capacidad global de discriminación del modelo logit completo dada el área bajo esta curva $AUC = 0.746$ alcanza un nivel de predicción aceptable según los umbrales de la Tabla 2 de Hosmer-Lemeshow.

```
> curvacompleto <- ROC(form=low~age+lwt+race+smoke+ptl+ht+ui+ftv,
 + plot="ROC",PV=T,MX=T,AUC=T,data=datos)
```

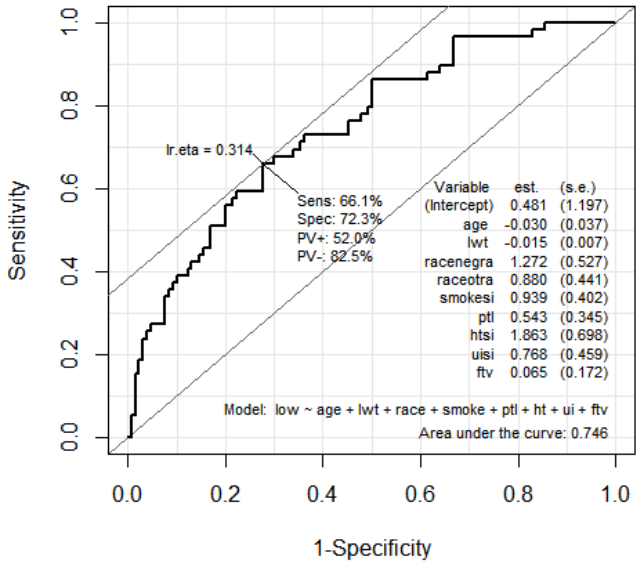

Figura 5: Gráfica de la curva ROC y punto de corte

Utilizando la librería *pROC* de R, también se obtienen las curvas ROC y se puede hacer fácilmente la comparación entre dos curvas. Por ejemplo, las curvas ROC para los modelos logísticos simple y completo (Figura 6) se obtienen como sigue:

```
> library("pROC")
> curva3 <- roc(modelo$y,modelo$fitted.values,ci=TRUE,plot=TRUE,
  + print.auc=TRUE,show.thres=TRUE)
> curva4 <- roc(modeloC$y,modeloC$fitted.values,ci=TRUE,plot=TRUE,
  + print.auc=TRUE,show.thres=TRUE)
> curva3; curva4
Call:
roc.default(response = modelo$y, predictor = modelo$fitted.values,
    ci = TRUE, plot = TRUE, print.auc = TRUE, show.thres = TRUE)
Data: modelo$fitted.values in 130 controls (modelo$y 0) < 59 cases (modelo$y 1).
Area under the curve: 0.6131
95% CI: 0.5245-0.7017 (DeLong)

Call:
roc.default(response = modeloC$y, predictor = modeloC$fitted.values,
     ci = TRUE, plot = TRUE, print.auc = TRUE, show.thres = TRUE)
Data: modeloC$fitted.values in 130 controls (modeloC$y 0) < 59 cases (modeloC$y 1).
Area under the curve: 0.7462
95% CI: 0.6721-0.8202 (DeLong)
```

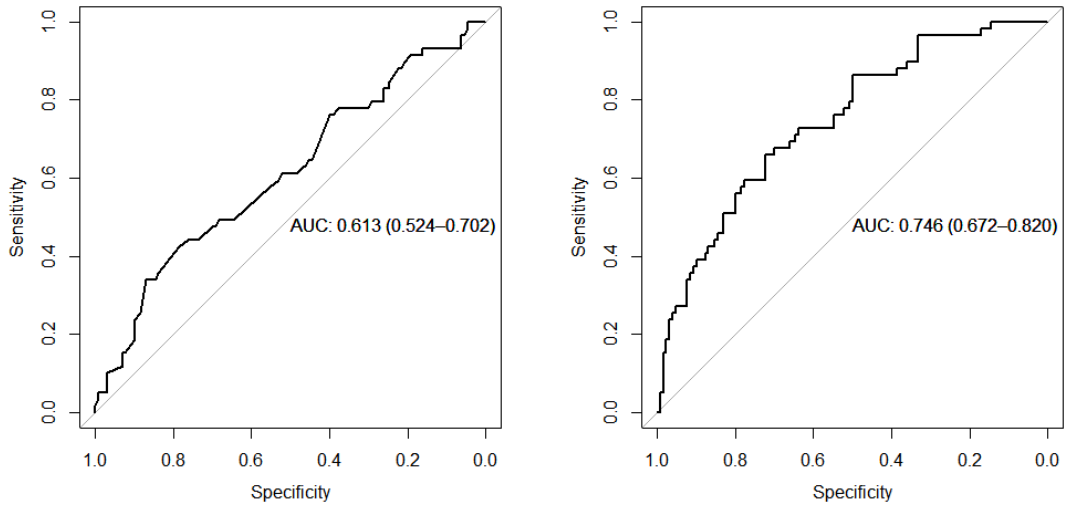

Figura 6: Gráficas de curvas ROC con sus áreas e intervalos

Este paquete de R incluye diferentes funciones para analizar una curva ROC de un clasificador y su comparación con la curva de otro modelo discriminador, tanto de forma global como parcial. Entre otras, permite representar la banda de confianza de la curva ROC (Figura 7) usando las siguientes funciones:

```
> ICsen <- ci.se(curva4,specificities=seq(0,1,.05))
> plot(ICsen,type="shape",col="lightblue")
```

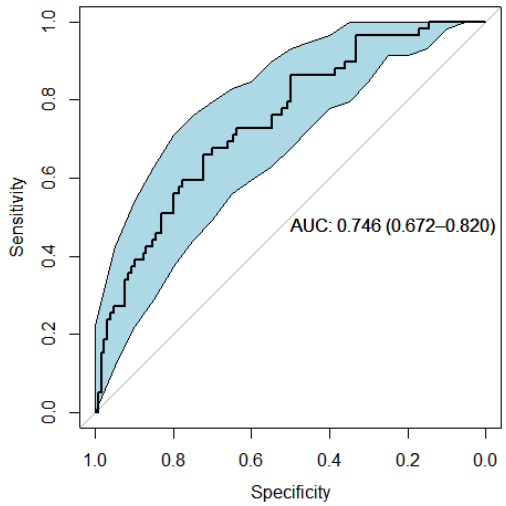

Figura 7: Curva ROC con banda de confianza para la sensibilidad

además de obtener los intervalos de confianza para el área bajo la curva se obtienen mediante la función *ci*, y la función *roc.test* realiza el contraste de comparación de ambas curvas:

```
> ci(curva3); ci(curva4)
    95% CI: 0.5245-0.7017 (DeLong)
    95% CI: 0.6721-0.8202 (DeLong)
> roc.test(curva3,curva4,reuse.auc=FALSE)
DeLong's test for two correlated ROC curves
data:  curva3 and curva4
```

```
Z = -2.8422, p-value = 0.00448
alternative hypothesis: true difference in AUC is not equal to 0
sample estimates:
AUC of roc1 AUC of roc2
  0.6131030   0.7461538
```

No obstante, se ha mencionado anteriormente, el modelo completo de regresión logística incluye covariables y factores no relevantes o influyentes; por ejemplo, hemos visto que la covariable fvt es una candidata para reducir el modelo de regresión logístico. Por lo que, debería repetirse el análisis del modelo de regresión logístico suprimiendo los términos no significativos, utilizando algún procedimiento de selección del modelo similar a los utilizados en la regresión lineal. En este sentido, como ejercicio ejecutar las siguientes instrucciones comentando las salidas de R:

```
> step(modelo0,scope=list(lower=modelo0,upper=modelo1),direction="forward")
> step(modelo1,scope=list(lower=modelo0,upper=modelo1),direction="backward")
> step(modelo0,scope=list(lower=modelo0,upper=modelo1),direction="both")
```

Por ejemplo, aquí se muestra el último paso de la salida con la selección en ambas direcciones (both):

```
Step:  AIC=217.99
low ~ ptl + lwt + ht + race + smoke + ui

        Df Deviance    AIC
<none>       201.99 217.99
- ptl    1   204.22 218.22
- ui     1   204.90 218.90
+ age    1   201.43 219.43
+ ftv    1   201.93 219.93
- smoke  1   207.73 221.73
- lwt    1   208.11 222.11
- race   2   210.31 222.31
- ht     1   209.46 223.46

Call:  glm(formula = low ~ ptl + lwt + ht + race + smoke + ui, family = binomial,
    data = datos)

Coefficients:
(Intercept)          ptl          lwt         htsi    racenegra     raceotra
   -0.08655      0.50321     -0.01591      1.85504      1.32572      0.89708
    smokesi         uisi
    0.93873      0.78570

Degrees of Freedom: 188 Total (i.e. Null);  181 Residual
Null Deviance:     234.7
Residual Deviance: 202  AIC: 218
```

Observar, que también puede aplicarse la función *stepAIC*, obteniendo el mismo modelo final y un breve resumen de los pasos desarrollados para seleccionar el modelo de regresión logística con términos relevantes:

```
> pasos <- stepAIC(modelo0,scope=list(lower=modelo0,upper=modelo1),direction="both")
> pasos$anova
Stepwise Model Path
Analysis of Deviance Table
```

```
Initial Model:
low ~ 1

Final Model:
low ~ ptl + lwt + ht + race + smoke + ui

     Step Df Deviance Resid. Df Resid. Dev      AIC
1                           188   234.6720 236.6720
2   + ptl  1 6.779384       187   227.8926 231.8926
3   + lwt  1 4.485746       186   223.4069 229.4069
4    + ht  1 7.443068       185   215.9638 223.9638
5  + race  2 5.113413       183   210.8504 222.8504
6 + smoke  1 5.952699       182   204.8977 218.8977
7    + ui  1 2.912098       181   201.9856 217.9856
```

Finalmente, se deja como ejercicio el análisis del modelo final seleccionado con 6 términos:

```
low ~ ptl + lwt + ht + race + smoke + ui
```

repitiendo los pasos utilizados en el ejemplo de modelo de regresión logística con dos factores, al igual que en el modelo completo realizado en este apartado. Así como la realización de la tabla resumen de las clasificaciones correctas e incorrectas realizadas a través de este modelo logit seleccionado, la curva ROC correspondiente junto con la medida de discriminación global y la determinación de un punto de corte óptimo, comparando la tabla de discriminación asociada a un punto de corte óptimo con la anterior.

Ejercicios prácticos

Ejercicio 4.2 *A partir de la base de datos birthwt sobre la clasificación en bajo peso o normal de los recién nacidos, analizar los datos observados a través de los siguientes apartados:*

(1) Crear un cuadro de datos declarando como factores la raza, fumadora, partos prematuros, hipertensión, irritabilidad uterina y visitas en el primer trimestre, considerando las variables raza si es blanca o no, pp si es 0 o no, y vp con niveles 0, 1, 2 o mayor.

(2) Analizar el modelo logístico completo con todos las dos covariables y los seis factores, teniendo en cuenta:

- *(a) Establecer el modelo de regresión logístico estimado para predecir un bebé de bajo peso.*
- *(b) Analizar los coeficientes y sus* odds-ratio*, determinando los intervalos de confianza y contrastes de significación.*
- *(c) Realizar el contraste de razón de verosimilitudes para la significación conjunta del modelo estimado.*
- *(d) Contrastar el ajuste de las probabilidades pronosticadas de bajo peso a través del modelo logit estimado a las probabilidades observadas en este experimento.*
- *(e) Determinar la capacidad global de discriminación del modelo logit completo y la tabla de clasificaciones correctas e incorrectas a través de un punto de corte óptimo.*
- *(f) ¿Qué diferencias principales encuentras entre este modelo logit completo y el modelo completo de regresión logística del ejemplo anterior?*

(3) Aplicar, si procede, un mecanismo de selección de términos relevantes en un modelo de regresión logística, y analizar el modelo logit resultante a través de los puntos indicados en el caso anterior.

Referencias

Bibliografía

- Everitt, B.S.; Hothorn, T. (2010). A Handbook of Statistical Analysis Using R. Chapman Hall.
- Faraway, J. (2004). Linear Models with R. CRC Press.
- Faraway, J. (2005). Extending the Linear Model with R. CRC Press.
- García Pérez, A. (2008). Estadística aplicada con R. UNED.
- Hosmer, D.W.; Lemeshow, S. (2000). Applied Logistic Regression. Wiley.
- Peña, D. (2002). Análisis de Datos Multivariantes. McGraw-Hill.

Recursos en Internet

- An Introduction to R: http://cran.r-project.org/doc/manuals/R-intro.pdf
- Comunidad R hispano: http://www.r-es.org/
- icebreakeR: http://www.ms.unimelb.edu.au/ andrewpr/r-users/icebreakeR.pdf
- Logistic regression (with R). Stanford NLP Group - C. Manning, 2007: http://nlp.stanford.edu/~manning/courses/ling289/logistic.pdf
- Practical Regression and ANOVA in R: on CRAN, Faraway, J.
- Quick-R: http://www.statmethods.net/
- Regresión y Análisis de la Varianza. Tusell, F. y Núñez, V. 2007: http://www.et.bs.ehu.es/etptupaf
- RStudio: http://rstudio.org/