Sección de Matemáticas
Universidad de La Laguna

Mariano Brage Escalona

Análisis de datos categóricos: regresión logística y multinomial

Categorical data analysis: logistic and multinomial regression

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Julio de 2020

Dirigido por
Enrique Gónzalez Dávila

Enrique Gónzalez Dávila
Departamento de Matemáticas,
Estadística e Investigación Operativa
Universidad de La Laguna
38200 La Laguna, Tenerife

Agradecimientos

Este trabajo fin de grado se lo dedico principalmente a mi familia, en especial a mi padre José Mariano Brage Camazano y a mi madre Manuela Escalona Otero, y también a mis amigos, en particular a Zebensuí Hernández Díaz y a Daniel Cao Labora quienes han sido un gran apoyo para la realización de este trabajo.

A mi tutor Enrique González Dávila por coger esta línea de trabajo y por su eterna paciencia durante la realización del trabajo. También quiero agradecer a las Doctoras Nieves Luisa González González y Erika Padrón de la Universidad de La Laguna y el Hospital Universitario de Canarias (HUC) por ceder los datos para la realización de este TFG.

Mariano Brage Escalona
La Laguna, 8 de julio de 2020

Resumen · Abstract

Resumen

En la actualidad el estudio de variables dependientes categóricas es útil en diversos ámbitos como en medicina, psicología, sociología, etc. Este trabajo desarrolla dos modelos de regresión sobre variable respuesta categórica, el primero, regresión logística, cuando se tiene una variable categórica dicotómica, y el segundo, regresión multinomial, cuando la variable es politómica. Estos dos modelos serán implementados usando datos reales ginecológicos con la intención de generar predicciones del estado de peso de los bebés a partir de información materna y variables obstétricas recogidas en la primera revisión a los tres meses de embarazo.

Palabras clave: *Regresión Logística – Regresión Multinomial – Dicotómica – Politómica.*

Abstract

Currently, the study of categorical dependent variables is useful in different fields such as health, psychology, sociology and so on. Two regression models on categorical variable answers will be developed in this essay, the first being the logistic regression model, which has one dichotomous variable, whilst the second is the multinomial model, in which the variable is polytomous. These two models will be implemented using real gynaecological data with the intention of generating predictions related to the weight of the babies from the maternal information and obstetrical variables gathered in the first medical examination in the first trimester of the pregnancy.

Keywords: *Logistic Regression – Multinomial Regression – Dichotomous – Polytomous.*

Contenido

Introducción

Actualmente existe una gran variedad de estudios de diferente índole en el que la variable de interés es de tipo categórica. Como por ejemplo, en medicina al examinar la presencia de una enfermedad en un paciente basándose en unas características recogidas, o bien, en Economía el estudio del grado de satisfacción de un cliente con el servicio contratado.

El modelo de regresión logística y el de regresión multinomial son técnicas analíticas que permiten relacionar una variable categórica con un conjunto de variables independientes que pueden ser categóricas o continuas para predecir sucesos. A partir de los coeficientes del modelo se puede interpretar los efectos que tienen estas variables sobre la respuesta. Esto puede ser muy útil en medicina para ayudar a diagnosticar a un paciente viendo que factores influyen más en la aparición de una enfermedad, o en Economía para ver que influye más en un cliente a la hora de valorar un servicio.

El Modelo Lineal Generalizado, MLG o en inglés General Linear Model, de aquí en adelante GLM, fue formulado por John Nelder y Robert Wedderburn para unificar los modelos de regresión lineal, logística y de Poisson. Esto ayuda a analizar variables, llamadas respuesta, que tengan una distribución de los errores distinta a la normal. Este modelo está formado por tres elementos, una función de distribución perteneciente a la familia exponencial (que puede ser una normal, una binomial, una poisson o una gamma), un predictor lineal formado por la combinación lineal de las variables independientes del modelo y la función de enlace, que proporciona la relación entre el predictor lineal y, la media de la función de distribución, entre las que se encuentran la función logit y y la función probit.

La regresión logística permite relacionar una variable dicotómica con un conjunto de variables independientes, que pueden ser dicotómicas, politómicas o continuas, para predecir sucesos. Es uno de los modelos que contiene los GLM, usando como función de enlace el logit, y ésta se considera una extensión de los modelos de regresión lineal, con la particularidad de que el recorrido de la función está

acotado en el intervalo $[0, 1]$ y, por otro lado, el procedimiento de estimación de los errores es el máximo-verosímil.

En ciertas aplicaciones hay que implementar el modelo de regresión multinomial, que es una generalización del anterior, en el que la variable respuesta es politómica.

Este trabajo se ha estructurado en tres capítulos. El primer capítulo introduce el modelo de regresión logística, relacionándolo con el modelo GLM, en el cuál se explicita la formulación, la interpretación de los coeficientes, la inferencia y la validación. Después en el segundo capítulo se presenta el modelo multinomial para variables respuesta politómicas. En el último capítulo se explica la implementación de los dos modelos en un problema real. El objetivo en los dos es conseguir predicciones del estado de peso de los bebés a partir de información materna (edad, peso, estatura, etc) y variables obstétricas (volumen de la placenta, índice de flujo, etc) recogida en la primera revisión a los tres meses de embarazo. En el modelo logístico se quiere predecir si el bebé es Pequeño para la Edad Gestacional (PEG) o no y en el multinomial predecir si éste es PEG, normal o Grande para la Edad Gestacional (GEG).

1

Regresión Logística

Se trata de una de las técnicas más conocidas y utilizadas para modelar una variable respuesta dicotómica en función de un conjunto de variables predictoras, que pueden ser continuas o categóricas. Este modelo es utilizado con frecuencia en medicina para modelar, por ejemplo, la probabilidad de que un paciente tenga una cierta enfermedad, en función de unas características recogidas sobre éste.

1.1. Nociones básicas

Antes de formular el modelo se necesita aclarar una serie de conceptos previos (Silva Ayçaguer y Barroso, 2004. [1]).

Definición 1. El *Odd* asociado a un suceso se define como la razón entre la probabilidad de que ese suceso tenga lugar y la probabilidad de que no ocurra.

$$Odd = \frac{p}{1-p}, \text{ siendo } p \text{ la probabilidad del suceso} \tag{1.1}$$

Si $p = 0 \Rightarrow Odd = 0$ y si $p = 1 \Rightarrow Odd = +\infty$. Es decir, el $Odd \in [0, +\infty)$. Si es conocido el *Odd* se puede calcular la probabilidad asociada simplemente despejando la p de la fórmula anterior 1.1.

$$p = \frac{Odd}{1 + Odd} \tag{1.2}$$

Definición 2. El Riesgo Relativo (RR) es la razón entre la probabilidad de que ocurra el suceso cumpliendo la condición A entre la probabilidad de que ocurra ese suceso cumpliendo la condición B.

$$RR = \frac{p_A}{p_B}$$

Un caso particular de la definición de RR es considerar la condición B como el complementario de A ($\bar{A}$).

En estudios de diagnóstico clínico y epidemiológico el riesgo relativo es uno de los objetivos principales a determinar, como por ejemplo, cuánto más probable es padecer el cáncer de pulmón si fumas que si no fumas, cuánto más probable es padecer una cierta enfermedad si una prueba diagnóstica es positiva frente a una negativa. En general, salvo que los estudios epidemiológicos realizados sean de cohorte o transversales, estas cantidades no son fácilmente obtenibles. El tipo de estudio más utilizado, estudios de casos y controles, no permiten la obtención del riesgo relativo, pero cuando la prevalencia de la enfermedad es relativamente baja es posible el cálculo de los odds ratio que se expone a continuación como un estimador del riesgo relativo.

Definición 3. El Odds ratio (OR) es la razón entre el *Odd* de un suceso bajo cierta condición A entre el *Odd* del mismo suceso bajo la condición complementaria.

$$OR = \frac{\frac{p_A}{1-p_A}}{\frac{p_{\bar{A}}}{1-p_{\bar{A}}}} \tag{1.3}$$

El OR en ocasiones se le denomina razón de productos cruzados, ya que su definición coincide con el cociente del producto de las cantidades observadas en la diagonal principal entre el producto de la diagonal secundaria cuando estas variables están en una tabla de contingencia.

Un OR igual a uno significa que las variables son independientes. Cuando el OR es mayor que uno entonces es más probable el suceso bajo la condición A. Mientras que uno menor que uno significa que es más probable el suceso bajo la condición $\bar{A}$.

Para poder entender mejor los modelos que se abordan en este trabajo se realiza una breve introducción de los modelos lineales generalizados.

1.2. Modelos Lineales Generalizados (GLM)

Para unificar los modelos de regresión lineal, logística y de Poisson (Agresti, 2007 [2]), aplica las mismas herramientas utilizadas en los modelos lineales generales a este tipo de variables, pero teniendo en cuenta que lo que se modeliza no es la variable observada sino su media o esperanza.

Sea la esperanza de Y condicionada por las observaciones de las variables dependientes, esto es, $E[Y|X_1 = x_1, ..., X_n = x_n] = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n$. Un modelo lineal generalizado está constituido por:

- La componente aleatoria correspondiente a la variable Y con función de distribución perteneciente a la familia exponencial (normal, log-normal, binomial, gamma o poisson).

- El predictor lineal, η_i, formado por la combinación lineal de las variables independientes X_i, $i = 1, \ldots\ n$.
- La función de vínculo. Esta función especifica la relación entre la esperanza condicionada y el predictor lineal. Hay diferentes funciones de enlace dependiendo del tipo de variable respuesta. Entre las más utilizadas está la identidad, la raíz cuadrada, el logaritmo natural, el log-log complementario, el probit y el logit.

El modelo de regresión lineal múltiple clásico considera la siguiente relación:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon_i \tag{1.4}$$

donde Y es la variable dependiente, α el intercepto, β_i con $i = 1, \ldots, k$ los coeficientes de regresión, X_i con $i = 1, \ldots, k$ las variables independientes y ϵ_i con $i = 1, \ldots, k$ los errores.

Los β_i miden el cambio promedio que se produce en Y para un incremento de una unidad en la j-ésima variable independiente, asumiendo que los valores del resto de variables no varían. Los errores debido al azar representan aquella variabilidad de Y atribuible a causas no controladas por el modelo lineal.

Este tipo de modelo presenta una serie de inconvenientes a la hora de aplicarlo a una variable dicotómica:

- La variable Y y los errores ϵ_i se distribuyen como una binomial en lugar de una normal.
- Por otro lado la varianza de Y no es constante, porque depende de la esperanza condicionada $E[Y|X = x] = p(x)$ y $Var[Y|X = x] = p(x)(1 - p(x))$.

Si se aplicase un modelo lineal sobre las probabilidades como variable dependiente, entonces los valores predichos podrían tomar valores mayores o menores que 1 y como se trata de una probabilidad esto no tendría sentido. Por esta razón se realiza una transformación, la llamada función de enlace *Logit(p)* como podemos observar en la figura 1.1 que corrige el problema ya que el dominio de esta función es $\mathbb{R}$.

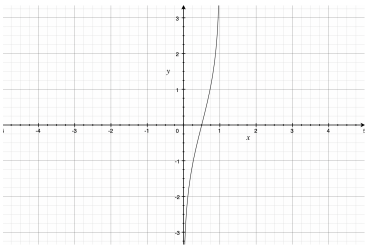


Figura 1.1: Función logit.

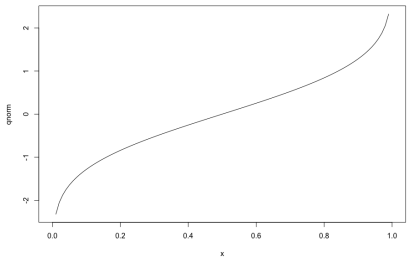


Figura 1.2: Función probit

Existen otras muchas transformaciones posibles, aunque destaca la función probit, definida como la imagen inversa de la función de distribución de la normal. (figura 1.2).

1.3. Construcción del modelo

Sea Z una variable de respuesta dicotómica con probabilidad de éxito p que ha sido codificada como 1 la categoría de interés y 0 la otra. Sean $X_1, ...X_n$ las variables explicativas que modelizan la probabilidad de éxito, tal que la $E[Y|X_1 = x_1, ..., X_n = x_n] = P[Y|X_1 = x_1, ..., X_n = x_n] = p(x_1, ..., x_n)$. El modelo de regresión logística permite modelizar una transformación de esta función $p(x_1, ...x_n)$ por medio de un predictor lineal tal como lo hace el modelo de regresión lineal. De esta manera utilizando la transformación $logit(p)$ se tiene:

$$ln\left(\frac{p}{1-p}\right) = logit(p) = \mu(x_1, ...x_n) = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n \tag{1.5}$$

A partir de este modelo, realizando unas operaciones algebraicas, se obtienen las probabilidades p:

$$p = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n}}{1+e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n}} \Leftrightarrow p = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}} \tag{1.6}$$

Si los coeficientes $\beta_i = 0$ con $i = 1, ..., n$ y $\alpha \neq 0$ entonces se tendría el modelo nulo (1.7), es decir, la variable Y es independiente de todas las variables explicativas. El coeficiente α es el $\ln(Odd)$ definido anteriormente.

$$logit(p) = \alpha \tag{1.7}$$

El coeficiente α sería el valor de la ventaja de respuesta de $Z = 1$ frente a $Z = 0$ cuando $\beta_i = 0$ o cuando $X_i = 0$, $\forall i = 1, ..., n$. El cociente de ventajas entre dos configuraciones de los valores de las variables explicativas, $x_1 = (1, x_{11}, ..., x_{1n})$ y $x_2 = (1, x_{21}, ..., x_{2n})$, es decir, el OR visto anteriormente en la definición 2, es:

$$OR(x_1, x_2) = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}} = \frac{e^{\alpha+\beta_1 x_{11}+...+\beta_n x_{1n}}}{e^{\alpha+\beta_1 x_{21}+...+\beta_n x_{2n}}} = e^{\beta_1(x_{11}-x_{21})+...+\beta_n(x_{1n}-x_{2n})} \tag{1.8}$$

Si x_1 y x_2 coinciden en todas las componentes salvo en la $k-ésima$ que se diferencia en una unidad se tendría $OR(x_1, x_2) = e^{\beta_k}$. Es decir, la exponencial del parámetro asociado a X_k es la cantidad por la que queda multiplicada la ventaja de respuesta $Z = 1$ frente a $Z = 0$ cuando el valor de X_k aumenta en una unidad, manteniendo fijos los valores de las demás variables.

1.4. Medidas de bondad de ajuste

1.4.1. Contraste de parámetros

Dada una muestra de n sujetos, un buen modelo debería asociar una alta probabilidad a los k sujetos en los que $Z = 1$ y una baja probabilidad a los $n - k$ sujetos en los que $Z = 0$. Para valorar este modelo se utiliza la verosimilitud del modelo, denotada por V y definida como:

$$V = \prod_{i=1}^{k} p_i^{Z_i} \prod_{i=k+1}^{n} (1 - p_i)^{1-Z_i} \tag{1.9}$$

Los mejores valores para los parámetros del modelo son aquellos que hacen que V sea lo más próxima a 1. Si $V = 1$ el modelo sería perfecto ya que asignaría $p_i = 1$ para los k sujetos y $p_i = 0$ para los $n - k$, como se observa en la ecuación 1.9.

Los coeficientes del modelo son calculados maximizando V, por ejemplo, con el método de Newton-Raphson. En el primer paso, se obtiene el valor de V_1, cuando se fija el primer parámetro del modelo como $a = \ln\left(\frac{\sum_{i=1}^{n} Z_i}{n - \sum_{i=1}^{n} Z_i}\right)$ y los demás $b_i = 0$, $\forall i = 1, \cdots, n$. Los valores finales de los parámetros que obtienen el máximo son denominados estimaciones máximo-verosímiles de los coeficientes α, β_1, $\cdots$, β_n. Por otra parte, la lejanía o *deviance* del modelo (L) queda definida como:

$$L = -2\ln(V) \tag{1.10}$$

Si $V = 1 \Rightarrow L = 0$, es decir el modelo sería perfecto, es decir, no cometería ningún error. Como no existe un modelo perfecto, en la práctica se tendrá que $V < 1$. Cuando V se aproxima a 0 se incrementa la lejanía y cuando se aproxima a 1, esta disminuye.

La lejanía del modelo nulo será denotada por L_0 y toma el mayor valor, ya que es un modelo simple al solo tener la constante. Esta sirve para evaluar si la incorporación de la variable es importante para el modelo.

Para evaluar la importancia de la inclusión de variables en el modelo se utiliza el cociente entre la verosimilitud del modelo sin la variable o variables que se quieren incluir y la verosimilitud del modelo con dichas variables, denominado razón de verosimilitudes (RV). RV se distribuye como una χ^2_m, siendo m el número de variables presentes en el modelo ampliado. Tiene m grados de libertad si se compara con el modelo nulo, si no tendrá $m - m^*$. [1]

$$RV = L_0 - L = -2\ln(\frac{V_0}{V}) \tag{1.11}$$

[1] m^* es el número de variables independientes que tiene el modelo con el que se compara.

Esta razón de verosimilitudes es muy útil para ver si hay diferencias significativas al incluir las variables, es decir, si tienen importancia real para que la variable respuesta Z tome el valor 1.

Otra forma de ver si una variable X_i es significativa o no para el modelo es la Prueba de Wald, que se trata de un test de hipótesis para cada uno de los coeficientes del modelo:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{cases} \tag{1.12}$$

El estadígrafo, Z_{Wald}, de esta prueba se obtiene dividiendo la estimación del coeficiente de interés por su error estándar, esto es:

$$Z_{Wald} = \frac{b}{se(b)} \tag{1.13}$$

Este estadístico sigue una distribución normal estándar. Esta prueba y la razón de verosimilitudes mencionada antes dan resultados similares en muestras grandes, pero esto no sucede cuando no son muestras muy grandes. En muestras no muy grandes estas pruebas producen diferentes resultados, en general se recomienda usar la RV (Silva Ayçaguer y Barroso, 2004. [1]).

1.4.2. Estadístico de Hosmer-Lemeshow

La evaluación del ajuste del modelo se realiza con los valores de Z_i y p_i las probabilidades calculadas a partir del modelo ajustado de la muestra de tamaño n. En primer lugar se ordenan los valores de p de menor a mayor, y se forman k grupos. Si $Z = 1$ se suman los valores de p dentro de cada uno de los grupos formados, estas sumas son los valores esperados E_i. Por último se cuentan los valores observados O_i en los que $Z = 1$. Por lo tanto, se realiza el siguiente contraste:

$$\begin{cases} H_0 : \text{El modelo ajusta bien los datos} \\ H_a : \text{Mal ajuste del modelo} \end{cases} \tag{1.14}$$

y el estadígrafo de Hosmer-Lemeshow se calcula de la siguiente manera (Silva Ayçaguer y Barroso, 2004. [1]).

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} + \sum_{i=1}^{k} \frac{(O_i^* - E_i^*)^2}{E_i^*} \tag{1.15}$$

donde $O_i^* = n_i - O_i$ y $E_i^* = n_i - E_i$ son los datos observados y esperados, respectivamente para los que $Z = 0$ y n_i es el grupo correspondiente, n_i es el tamaño del grupo y k es el grupo en cuestión. Sigue una distribución χ^2 con $k - 2$ grados de libertad siendo k el número de grupos formados (Agresti, A.(2007)[2]). La hipótesis nula establece que no hay diferencias entre los valores observados y los pronosticados, es decir, el modelo ajusta bien los datos. Por tanto, si el ajuste es bueno, se espera un valor alto de p-valor, superior o igual a 0,05.

1.4.3. Medidas tipo R^2

En regresión lineal se utiliza el coeficiente de determinación, llamado R^2, que se define como la proporción de la varianza total de la variable explicada por la regresión y refleja la bondad de ajuste de un modelo. Este coeficiente oscila entre 0 y 1, cuanto más próximo a 1 sea su valor, mayor será el ajuste del modelo, y cuanto más próximo a 0 menos ajustado estará el modelo.

$$R^2 = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2} \tag{1.16}$$

En el caso de la regresión logística se podría considerar esta medida substituyendo $\hat{y}_i$ por $n\hat{p}_i$, siendo $\hat{p}_i$ la probabilidad estimada. Los estimadores verosímiles no son los que maximizan esta medida ni tiene en cuenta la dependencia de la varianza de Y respecto de p. Por todo esto se utilizan unas medidas alternativas, denominadas pseudo R^2, las principales son:

- R^2 de McFadden: compara la distancia del modelo ajustado, L, con la del modelo nulo, L_0. Proporciona una idea de cuánto se reduce la distancia de los datos al ajustar el modelo. Su valor teóricamente oscila entre 0 y 1, pero rara vez se aproxima a 1. Se considera un buen ajuste cuando $R^2_{MF} \in [0,2; 0,4]$ y excelente para valores más grandes.

$$R^2_{MF} = 1 - \frac{L}{L_0} \tag{1.17}$$

- R^2 de Cox-Snell: este coeficiente se obtiene en función de la verosimilitud aunque se puede expresar también en función de las distancias. El valor de $R^2_{CS} \in [0; 1 - \left(\sqrt[N]{V_0}\right)^2]$ y queda definido como:

$$R^2_{CS} = 1 - \left(\frac{V_0}{V}\right)^{\frac{2}{N}} \tag{1.18}$$

  donde V es la verosimilitud y N es el tamaño de la muestra.
- R^2 de Nagelkerke: se trata de una modificación del índice de Cox-Snell. Este coeficiente $R^2_N \in [0; 1]$ por lo que se puede interpretar igual que el coeficiente de regresión R^2, aunque es más difícil que tome valores próximos a 1.

$$R^2_N = \frac{R^2_{CS}}{1 - V_0^{\frac{2}{N}}} \tag{1.19}$$

1.4.4. Curvas ROC

Las curvas ROC, acrónimo de Relative Operating Characteristic, o en español COR (Característica Operativa del Receptor), se comenzaron a utilizar durante la II Guerra Mundial para el análisis de señales de radar. En ese momento fueron parte de lo que se conoce como Teoría de Detección de Señales (Green y Swets, 1966. [3]). La utilización de las curvas ROC se ha extendido en diferentes ámbitos, y en particular, en el campo de la medicina (Del Valle Benavides, 2017. [6]). En esta última, el análisis ROC se utiliza de manera muy extensa en problemas relacionados con la epidemiología. También en radiología se utiliza para evaluar nuevas técnicas de diagnóstico por imagen. Hace poco se han mostrado muy útiles para la evaluación de técnicas de aprendizaje automático. La primera aplicación en este ámbito fue hecha por Spackman que demostró su valor para la comparación de diferentes algoritmos de clasificación.

Supóngase que se tiene una variable aleatoria Z que se distribuye como una Bernoulli de parámetro p tal que $p = P_r(Z = 1)$ es la probabilidad de que el individuo esté enfermo. Sea X una variable aleatoria que mide cierta característica en cada individuo, cuyo resultado puede ser continuo o discreto. Sea c el valor umbral o punto de corte que se utiliza para construir la variable Y, la dicotomización de X, como se muestra en la ecuación 1.20.

$$Y = \begin{cases} 1 \text{ si x} \geq \text{c el paciente da positivo} \\ 0 \text{ si x} < \text{c el paciente da negativo} \end{cases} \tag{1.20}$$

Tabla 1.1: Tabla de clasificación

	$Z = 1$ (Enfermo)	$Z = 0$ (Sano)
$Y = 1$ (Prueba+)	Verdadero positivo(V_+)	Falsos positivo(F_+)
$Y = 0$ (Prueba-)	Falso negativo(F_-)	Verdadero negativo(V_-)

Para poder construir la curva ROC se necesita definir antes la sensibilidad y la especificidad.

Definición 4. La sensibilidad (S) es la probabilidad de, dado un individuo enfermo, que la prueba lo clasifique como enfermo.

$$S = P_r(Y = 1|Z = 1) = \frac{P_r(Y = 1 \cap Z = 1)}{P_r(Z = 1)} \tag{1.21}$$

Definición 5. La especificidad (E) es la probabilidad de, dado un individuo sano, que la prueba lo clasifique como sano.

$$E = P_r(Y = 0|Z = 0) = \frac{P_r(Y = 0 \cap Z = 0)}{P_r(Z = 0)} \tag{1.22}$$

En la práctica, la sensibilidad se estima por medio de la proporción de individuos que presentan el evento de interés y que son clasificados por la prueba como portadores de dicho evento y la especificidad por la proporción de individuos que no lo presentan y son clasificados por la prueba como tal. La curva ROC es un gráfico donde se representa la sensibilidad frente a 1 - especificidad, falsos positivos, para cada posible valor umbral o punto de corte en la escala de resultados de la prueba en estudio.

$$S = \frac{V_+}{V_+ + F_-} \; y \; E = \frac{V_-}{V_- + F_+} \tag{1.23}$$

Como se tienen probabilidades en los dos ejes la curva queda definida en el cuadrado $[0, 1] \times [0, 1]$.

El área bajo la curva (AUC) es el estadístico por excelencia para medir la capacidad discriminante de la prueba. También se utiliza para comparar pruebas entre sí y ver cuál es mejor. Dicha área se calcula de la siguiente forma:

$$AUC = \int_0^1 ROC(t)d(t) \tag{1.24}$$

Según Hosmer & Lemeshow (2000) [5] el valor de AUC se interpreta de la siguiente manera:

- Si AUC$= 0,5$, no hay discriminación, es como la analogía de tirar una moneda.
- Si AUC$\in [0,5; 0,7)$, hay poca discriminación.
- Si AUC $\in [0,7; 0,8)$ es una discriminación aceptable.
- Si AUC $\in [0,8; 0,9)$ es una discriminación excelente.
- Si AUC$\geq 0,9$ es una discriminación casi perfecta.

El Índice de Youden (YI), formulado por William J. Youden en 1950, es otro estadístico para estudiar la capacidad discriminante de la prueba. Este índice maximiza de forma conjunta la sensibilidad y la especificidad y se define como:

$$YI = \text{máx}(S + E - 1) \tag{1.25}$$

Cada punto de corte c está asociado a un valor de YI, y por tanto, puede servir para cuantificar la decisión y elegir el punto de corte óptimo.

En una curva ROC, el YI es la distancia vertical máxima desde la curva ROC a la diagonal principal, siendo c_{YI} el punto de corte óptimo.

Si tenemos dos curvas ROC, es mejor en el sentido de discriminante, aquella que tenga mayor área. Si éstas fuesen iguales, podría ser que fueran la misma curva o porque una ofrece una prueba más sensible y otra más específica. Una vez

dibujada toca escoger el punto de corte o valor umbral, lo ideal sería que tuviera una alta sensibilidad y especificidad. Esto puede hacerse de diferentes maneras dependiendo de si se quiere maximizar la sensibilidad y especificidad de forma conjunta o solamente una de ellas. Para el primer caso se escoge como valor de c el que mayor índice de Youden tenga, si en cambio se quiere maximizar la sensibilidad el valor de c sería el punto más cercano al vértice (0,1) y por último si se quiere maximizar la especificidad se escoge como valor de c el que minimice los costes de los resultados erróneos. En la figura 1.3 podemos ver un ejemplo de una curva ROC generada con R.

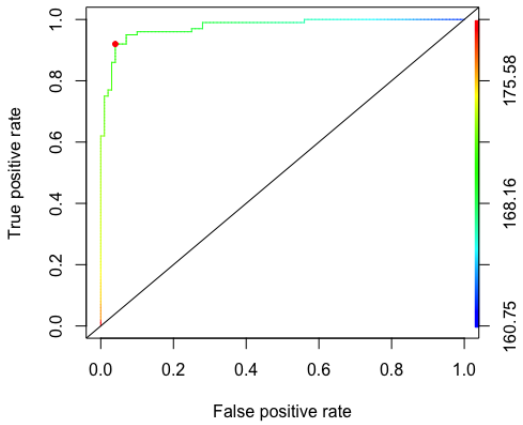


Figura 1.3: Curva ROC

1.5. Modelización

Tanto el test de la razón de verosimilitud como el test de Wald son instrumentos a utilizar para llevar a cabo el proceso de construir un modelo de regresión logística a partir de una base de datos. Hay diferentes estrategias estrategias para llevarlo a cabo.

- *Selección hacia adelante*: en cada etapa se añade la mejor variable independiente aún no seleccionada.
- *Selección hacia atrás*: parte con un modelo con todas las variables predictoras y en cada etapa se va eliminando la peor variable independiente hasta obtener un modelo en el que todas las variables predictoras son significativas.
- *Selección paso a paso*: combina las dos estrategias anteriores.

1.6. Residuos

Una vez se tiene un modelo de regresión logística que se ajuste a los datos de la muestra hay que analizar si el modelo cumple el supuesto de linealidad entre el $logit(p)$ y las variables independientes $X_1, \cdots, X_n$. Sea H la matriz de predicción o matriz *hat* definida de la siguiente forma:

$$H = V^{\frac{1}{2}} X (X'VX) X' V^{\frac{1}{2}}, \tag{1.26}$$

siendo X la matriz de predicción y $V = diag(\hat{p}(1-\hat{p}))$. Sea h_i el $i-ésimo$ elemento de la diagonal de H, con $h_i \in (0, 1)$ con un valor medio de p/n.Un residuo es una medida que expresa la diferencia entre las respuestas observadas y predichas por el modelo (Agresti, 2007.[2]). Se considera un valor atípico aquel cuyo residuo en valor absoluto es mayor que 2.

- residuos de Pearson.

$$r_i = \frac{y_i - n_i p_i}{\sqrt{n_i p_i (1 - p_i)}} \tag{1.27}$$

- residuo de Pearson estandarizado.

$$r_{si} = \frac{r_i}{\sqrt{1 - h_i}} \tag{1.28}$$

- residuos de la lejanía o raíz cuadrada de la contribución de cada observación a la lejanía, d_i.

$$d_i = \begin{cases} -\sqrt{-2\ln(p_i)} \text{ si } \mathrm{y_i} = 1 \\ \\ -\sqrt{-2\ln(1 - p_i)} \text{ si } \mathrm{y_i} = 0 \end{cases} \tag{1.29}$$

- residuos de la lejanía estandarizados.

$$d_{si} = \frac{d_i}{\sqrt{1 - h_i}} \tag{1.30}$$

1.7. Outliers

Al igual que en regresión lineal, algunas observaciones pueden influir mucho en la determinación de los parámetros estimados. Un outlier o valor atípico es una observación que dista numéricamente del resto de los datos. Los valores atípicos son en ocasiones una cuestión subjetiva y existen numerosos métodos para clasificarlos,

- De Cook: análogo, en regresión logística, al estadístico de influencia de Cook. Evalúa cuánto cambiarían los residuos al excluir un individuo en el cálculo de los coeficientes del modelo. Si es mayor que 1 la observación se considera influyente.

$$\Delta B_j = \frac{r_{si}^2 h_j}{1 - h_j} \tag{1.31}$$

- Valor de influencia o leverage: evalúa la influencia relativa de una observación en el ajuste del modelo. El leverage para la observación i-ésima es el elemento i-ésimo de la diagonal principal de la matriz H. Se considera un punto de alta influencia aquel cuyo valor $h_i > \frac{2p}{n}$.
- DfBetas: evalúa la diferencia en los coeficientes del modelo de regresión que resulta de la exclusión de un caso particular. Se calcula un valor para cada término del modelo, incluyendo la constante. Permiten ver sobre qué variable del modelo es influyente cada observación.

$$Dfbeta\beta_i^{(j)} = \beta_i - \beta_i^{(j)}, \tag{1.32}$$

donde β_i es el valor del coeficiente cuando todos los casos están incluidos y $\beta_i^{(j)}$ es el valor del coeficiente cuando el j-ésimo caso es excluido.

2

Regresión Multinomial

La regresión multinomial es una generalización del modelo de regresión logística en el que la variable dependiente, Y, tiene tres o más categorías. Este modelo asume que Y tiene una distribución multinomial (Agresti, 2007 [2] Silva Ayçaguer y Barroso, 2004 [1]).

2.1. Formulación del modelo

Sea Y una variable respuesta categórica con J categorías y sean π_1, π_2, ..., π_J las probabilidades asociadas, tal que $\sum_{j=1}^{J} \pi_j = 1$. Este modelo se construye tomando como respuesta base una de las categorías, por ejemplo, la última, J, y se define un modelo logit con respecto a ella:

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \sum_{k=1}^{K} \beta_{jk} X_{jk},\ j = 1, \ldots, J-1 \tag{2.1}$$

El modelo tiene $J-1$ ecuaciones con sus propios parámetros. Cada uno de estos parámetros expresa el efecto con respecto a la categoría de referencia. Cuando $J = 2$ se simplifica a una única ecuación, obteniendo el modelo logístico explicado en el capítulo 1. Utilizando la ecuación 2.1 y fijando dos categorías cualesquiera, a y b, se tiene que:

$$\begin{aligned}
\log\left(\frac{\pi_a}{\pi_b}\right) &= \log\left(\frac{\frac{\pi_a}{\pi_J}}{\frac{\pi_b}{\pi_J}}\right) \\
&= \log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right) \\
&= \alpha_a + \beta_{1a} X_1 + \cdots + \beta_{ka} X_K - \alpha_b - \beta_{1b} X_1 - \cdots - \beta_{kb} X_K \\
&= (\alpha_a - \alpha_b) + (\beta_{1a} - \beta_{1b}) X_1 + \cdots + (\beta_{Ka} - \beta_{kb}) X_K
\end{aligned} \tag{2.2}$$

Así se obtiene la ecuación logit de la categoría a con respecto a una categoría b cualquiera, $\alpha = \alpha_a - \alpha_b$, $\beta_1 = \beta_{1a} - \beta_{1b}$, ... y $\beta_k = \beta_{ka} - \beta_{kb}$. Los programas estadísticos ajustan simultáneamente todas las ecuaciones y, los errores que se obtienen para los parámetros estimados son más pequeños que los de regresión. Los parámetros del modelo se estiman utilizando el método de máxima verosimilitud como en el caso de regresión logística. Después de calcular estos parámetros se pueden calcular las probabilidades predichas de cada categoría despejando π_j de la ecuación 2.1.

$$\hat{\pi_j} = \frac{e^{\alpha_{ji} + \sum_{i=1}^{n} \beta_j X_i}}{1 + \sum_{h=1}^{J-1} e^{\alpha_h + \sum_{i=1}^{n} \beta_{hi} X_i}}, \; j = 1, \ldots, J \tag{2.3}$$

Un caso particular de la regresión multinomial es la regresión ordinal que se utiliza cuando las categorías de la variable respuesta tienen orden y cumplen la prueba de líneas paralelas, es decir, que el comportamiento de las variables independientes en cada categoría de la variable Y es igual. Esta regresión proporciona modelos más sencillos de interpretar que el modelo multinomial (2.1). Esto se debe a que la solución de los coeficientes β_i es la misma para todas las ecuaciones y la única diferencia es el coeficiente α.

El modelo de regresión ordinal, a diferencia del multinomial, acumula las probabilidades de las categorías anteriores y no utiliza la última categoría como referencia pues la probabilidad acumulada es igual a 1. De esta forma el modelo de regresión ordinal puede ser escrito como:

$$logit(P(Y \leq j)) = \ln\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \ln\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right) \tag{2.4}$$

2.2. Medidas de bondad de ajuste

2.2.1. Contraste sobre los parámetros

Dada una muestra de tamaño n de las variables Y_{ji} y X_k con $i = 1, \ldots, n$ tamaño de la muestra, $j = 1, \ldots, J$ número de categorías de Y y $k = 1, \ldots, K$ número de variables regresoras X, se puede definir en función de éstas los valores de Z y π. La función de verosimilitud del modelo V queda definida como:

$$V = \prod_{i=1}^{n} p_{1i}^{Y_{1i}} \ldots p_{Ji}^{1 - \sum_{j=1}^{J-1} \sum_{i=1}^{n} Y_{ji}} \tag{2.5}$$

De igual modo que en regresión logística se busca maximizar la verosimilitud, o de forma equivalente minimizar la distancia L:

$$L = -2\ln(V) \tag{2.6}$$

Para evaluar la significación de cada una de las variables principalmente se utilizan el estadístico de Wald y el estadístico condicional de razón de verosimilitud. Es decir, se realiza el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \beta_{jk} = 0 \\ H_a : \beta_{jk} \neq 0 \end{cases} \tag{2.7}$$

Como en el capítulo anterior,utilizando el estadíatico de Wald (1.13) que sigue una distribución χ^2_1, se puede contrastar si cada uno de los estimadores de los parámetros son significativamente distintos de 0. Una variable se considera significativa si $Z_{Wald} > \chi^2_{1;\alpha}$, es decir, se rechaza la hipótesis nula.

2.2.2. Significación de cada variable regresora

Este contraste de hipótesis estudia que ocurre en el modelo si se elimina una variable regresora del modelo, es decir, se evalua si los coeficientes que acompañan a dicha variable son nulos. Para ello se calcula la distancia del modelo eliminando la variable regresora X_i, L_{-i}. Realizando la diferencia entre las distancias del modelo sin la correspondiente variable X_i y el modelo final, L_f, obteniendo un estadístico que se distribuye como una $\chi^2_{(k-1)(J-1)}$, siendo k el número de variables regresoras. Se rechaza la hipótesis nula si $p(\chi^2_2 > L_{-i} - L_f)$.

2.2.3. Contraste de bondad de ajuste del modelo

Se puede probar que ninguna variable del modelo es significativa utilizando la razón de verosimilitudes que se utilizó en el capítulo anterior, que consiste en la diferencia entre las distancias inicial y final. Este estadístico sigue una distribución $\chi^2_{k(J-1)}$, siendo k el número de variables regresoras del modelo. El p-valor de este test vendrá dado por $p(\chi^2_{k(J-1)} > L_0 - L_f)$, siendo k el número de variables regresoras. Con esta prueba se estudia si todos los coeficientes β son nulos.

2.2.4. Medidas tipo R^2

La calidad de ajuste del modelo multinomial se puede medir utilizando los coeficientes de determinación pseudo R^2 ya definidos en el capítulo 1. Se utilizan los pseudo R^2_{MF}, R^2_{CS} y R^2_N. Para comparar modelos multinomiales con diferente número de variables regresoras se introducen coeficientes pseudo R^2 ajustados. El más conocido es el de Mc-Fadden (Pando y San Martín (2004) [8]), definido como:

$$Adj - R^2_{MF} = 1 - \frac{0,5\dot{L}_f + k + 1}{0,5\dot{L}_0 + 1}, \text{ siendo } k \text{ el número de variables regresoras.} \tag{2.8}$$

2.2.5. Tasa de clasificaciones correctas

Otra forma de cuantificar la bondad de ajuste del modelo consiste en comparar los resultados obtenidos por este frente a los observados, obteniendo así una matriz de clasificación. De forma que el número de individuos de cada categoría bien clasificados se contabiliza en la diagonal de esta matriz. Calculando la traza de la matriz dividiendo por el número total y multiplicando por 100 se obtiene el porcentaje de aciertos o también llamado tasa de clasificaciones correctas.

3

Aplicaciones de los modelos de regresión logística y regresión multinomial

En este capítulo se van a aplicarlos modelos introducidos en los dos capítulos anteriores a datos del campo de la medicina, y más concretamente en el campo de la obstetricia y ginecología. En primer lugar se aplicará el modelo de regresión logística con la idea de detectar los recién nacidos con peso bajo a partir de información del primer trimestre de embarazo, y el segundo ejemplo, utilizará regresión multinomial, con la idea de detectar tanto los recién nacidos con peso bajo como como exceso de peso.

Un recién nacido es considerado pequeño para la edad gestacional (PEG) si su peso al nacer está por debajo del percentil 10 de los pesos de recién nacidos de la población normal. Las medidas morfométricas maternas como la edad, la estatura y el peso de la madre, y las medidas biométricas fetales son factores importantes a la hora de diagnósticar con antelación si un recién nacido será PEG. No corregir este problema puede afectar tanto al feto, al parto, así como al neonato. Entre estos destaca: pérdida de bienestar fetal en el parto, aumento del número de cesáreas, aumento de morbimortalidad, depresión perinatal ante un parto mal tolerado, hipoglucemia, hiperviscosidad, aspiración de meconio, hipertensión pulmonar persistente, etc. (Gómez-Roig, 2012 [9]).

Un recién nacido es grande para la edad gestacional (GEG) si su peso está por encima del percentil 90 de los pesos de los recién nacidos de población normal. Los factores maternos de riesgo que pueden afectar principalmente son: peso, talla, edad (si supera los 35 años), diabetes pregestacional, multiparidad y tamaño uterino y placentario. Las consecuencias de no prevenirlo pueden afectar tanto a la madre, al parto, como al recién nacido. Provocándole una cesárea, desgarros en el canal del parto, hemorragias, complicaciones anestésicas y quirúrgicas e infecciones. Al bebé podría ocasionarle durante la gestación la muerte, una miocardiopatía, malformaciones congénitas, traumatismos obstétricos como la distocia de hombros y la lesión del plexo braquial y hemorragia subgaleal y cefalohematoma o una vez nacido como la asfixia, el síndrome de aspiración de meconio, una hipertensión pulmonar persistente, etc. (Unceta-Barrenechea et al. 2008 [10]).

El objetivo del primer ejemplo es conseguir predecir con un modelo de regresión logística si el bebé va a ser PEG, utilizando los datos de los tres primeros meses con la finalidad de tomar medidas durante la gestación y evitar problemas al recién nacido. Y por último, se añade una categoría más a la variable respuesta para predecir si el recién nacido será PEG, adecuado para su edad gestacional (AEG) o GEG y poder así, emplear el modelo de regresión multinomial.

En este capítulo se utilizan los datos de 1.000 madres gestantes para aplicar los modelos de regresión logística y multinomial, estos datos han sido proporcionados por las Doctoras Nieves Luisa González González y Erika Padrón de la Universidad de la Laguna y el Hospital Universitario de Canarias (HUC).

3.1. Descripción de la base de datos

Del total de partos incluidos en la base de datos, 988 son los que llegaron a término. Los datos del fichero se agrupan en diferentes bloques: información inicial de la madre (como la edad, el peso, la estatura, etc ...), información del parto anterior (como el tipo de parto previo, si el parto previo fue prematuro e hipertensión gestacional en el parto anterior), información ecográfica del 1º trimestre (recogida mediante diversas pruebas sobre la placenta y el feto), información final de la gestación (como el peso al final del embarazo) e información del bebé (peso al nacer) (ver tabla 3.1).

Tabla 3.1: Identificación de variables utilizadas.

Bloques	**Nombre**	**Descripción**
Información inicial de la madre	Edad materna	Edad de la madre en años.
	Peso materno	Peso de la madre en Kg.
	Estatura	Estatura de la madre en cm
	IMC	Indice de Masa Corporal calculado a partir del cociente del peso y la estatura de la madre al cuadrado.
	Etnia	1=Negro, 2=Mezcla, 3=Oriental y 4=Caucásico.
	Concepción	1=Espontáneo, 2=Ovulación y 3=Inseminación.
	Hipertensión crónica	1=sí y 0=no.
	Nulípara	1=sí y 0=no.
	Obesidad	Calculada a partir del IMC que toma los valores 1=sí y 0=no.
	Diabetes pregestacional	1=sí y 0=no.

Bloques	**Nombre**	**Descripción**
	Años de diabético	Duración de la diabetes en años.
	Fumadora	1=sí y 0=no.
Información del parto anterior	Parto anterior	1=sí 0=no.
	Parto prematuro	1=sí y 0=no.
	Hipertensión gestacional	1=sí y 0=no.
Información ecográfica del 1 trimestre	Vol	Volumen de la placenta.
	VI	Índice de vascularización de la placenta.
	FI	Índice de flujo de la placenta.
	VFI	Índice de vascularización-flujo de la placenta.
	CRL	Longitud craneal del bebé.
	NT	Longitud de la translucencia nucal del bebé.
	DeltaNT	Índice de la longitud de la translucencia nucal del bebé.
	UtPI	Índice de pulsatilidad de la arteria uterina.
	PbhCGMoM	Proteína de la gonadotropina coriónica beta-humana.
	PAPPA-MoM)	Proteína a-plasmática asociada al embarazo.
Información durante la gestación	Diabetes gestacional	1=sí y 0=no.
	Hipertensión gestacional	1=sí y 0=no.
	Preeclampsia	1=sí y 0=no.
	Edad gestacional	Duración de la gestación en días.
	Tipo de parto	1=Vaginal y 0=cesárea.
	Peso final	Peso de la madre al final en Kg.
	Incremento de peso	Diferencia de peso de la madre entre el comienzo y el final en Kg.
Información del bebé	Vivo	1=sí y 0=no.
	Peso	Peso del bebé en gramos.
	Percentil peso	
	Malformación	1=sí y 0=no.
	Sexo	-1=femenino y 1=masculino.
	PEG	1 = Pequeño para la edad gestacional y 0 = no

Una descripción de las principales variables recogidas, en función de la clasificación del peso del recién nacido es mostrada en la tabla 3.2. La información de las variables del bloque de parto anterior están referidas a las madres multíparas. La tabla 3.2 proporciona información sobre la media y la desviación típica en el caso de las variables continuas y el número de casos y sus frecuencias.

Tabla 3.2: Estadística descriptiva en función de las categorías de peso del recién nacido.

Variables	**PEG n=263**	**AEG n=612**	**GEG n=113**	**p-valor**
Edad materna	29,7 (6,6)	31,3 (5,5)	32,2 (5,7)	<0,001
Peso materno	66,2 (14,6)	69,9 (14,6)	76,2 (15,7)	<0,001
Estatura	160,4 (6,3)	163 (6,1)	164.1 (5,7)	<0,001
IMC	25,7 (5,5)	26,3 (5,2)	28,3 (5,6)	<0,001
Caucásico	258 (98 %)	604 (99 %)	111 (98 %)	0,4601
Concepción (espontáneo)	254 (97 %)	584 (95 %)	107 (95 %)	0,5231
Hipertensión crónica	9 (3 %)	18 (3 %)	2 (3 %)	0,9013
Nulípara	183 (70 %)	351 (57 %)	48 (43 %)	<0,001
Obesidad	47 (18 %)	135 (22 %)	40 (35 %)	<0,001
Diabetes pregestacional	4 (1,5 %)	38 (6 %)	27 (24 %)	<0,001
Años con diabetes	5,8 (3,4)	10,62 (7,6)	9,2 (9)	<0,001
Fuma	84 (32 %)	115 (19 %)	15 (13 %)	<0,001
Parto anterior				
Tipo de parto	67 (84 %)	237 (91 %)	51 (79 %)	0,0146
Prematuro	19 (27 %)	25 (11 %)	10 (19 %)	0,0018
Hipertensión gestacional	9 (14 %)	18 (12 %)	9 (24 %)	0,1294
Volumen(Vol)	52,1 (17,5)	64,8 (20,3)	74,3 (23,2)	<0,001
VI	8,6 (5,1)	9,5 (4,5)	9,3 (3,8)	<0,001
FI	47,6 (4,4)	48,7 (4,6)	49,5 (4,7)	<0,001
VFI	4,2 (2,8)	4,7 (2,5)	4,7 (2,1)	<0,001
CRL	61,5 (7,6)	64,5 (7,7)	66,4 (7,7)	<0,001
NT	1,6 (0,4)	1,8 (0,4)	1,9 (0,4)	<0,001
DeltaNT	0,02 (0,4)	0,07 (0,4)	0,12 (0,4)	<0,001
UtPI	2,1 (0,7)	1,8 (0,6)	1,7 (0,4)	<0,001
bhCGMoM	1,2 (0,9)	1,3 (0,9)	1,2 (0,7)	<0,001
PAPPA-MoM	1 (0,6)	1,2 (0,7)	1,3 (0,6)	<0,001
Diabetes gestacional	30 (11 %)	134 (22 %)	25 (22 %)	<0,001
Hipertensión gestacional	3 (1 %)	6 (1 %)	2 (2 %)	0,7624
Preeclampsia	34 (13 %)	31 (5 %)	8 (7 %)	<0,001
Edad gestacional	273 (21,1)	276,5 (13,2)	275,5 (12,4)	<0,001
Tipo de parto	210 (80 %)	520 (85 %)	80 (71 %)	<0,001
Peso final	75,4 (14,2)	80,2 (14,1)	88,3 (15,8)	<0,001
Incremento de peso	9,2 (4,9)	10,2 (5,1)	12,1 (4,9)	<0,001
Recién nacido				
Nacido vivo	253 (96 %)	610 (99,7 %)	112 (99 %)	<0,001
Peso (g.)	2508,3 (501,3)	3241,9 (390,7)	4085,7 (452,1)	<0,001
Percentil peso	3,7 (3,1)	45,8 (23,4)	96,7 (3,3)	<0,001
Malformación	4 (2 %)	2 (0,3 %)	1 (1 %)	0,1507
Sexo (masculino)	116 (44 %)	313 (51 %)	65 (58 %)	0,0382

3.2. Detección de recién nacidos con peso bajo

El objetivo en esta sección es encontrar un modelo que permita predecir si un feto va a ser PEG utilizando la información obtenida en los tres primeros meses de embarazo. En primer lugar se construye un modelo utilizando las variables edad, peso, estatura, concepción, hipertensión crónica, nulípara, obesidad, diabetes pregestacional, Vol, VI, FI, VFI, DeltaNT, UtPI, bhCGMoM y PAPPAMoM. Este modelo será identificado como *intro*. En este modelo no se tendrán en cuenta la variable Etnia, ya que el 99 % de las madres son caucásicas, ni las variables del bloque de información del parto anterior. En la tabla 3.3 se muestra el estudio descriptivo en función de si es PEG o no.

Tabla 3.3: Estadística descriptiva en función de si el recién nacido es pequeño o no.

Variables	**PEG n=263**	**no PEG n=725**	**p-valor**	**Total n=988**
Edad materna	29,7 (6,6)	31,5 (5,6)	<0,001	31,0 (5,9)
Peso materno	66,2 (14,6)	70,9 (14,9)	<0,001	69,6 (15,0)
Estatura	160,4 (6,3)	163,2 (6,0)	<0,001	162,4 (6,2)
IMC	25,7 (5,5)	26,6 (5,3)	<0,001	26,4 (5,4)
Caucásico	258 (98 %)	715 (99 %)	0,309	973 (99 %)
Concepción (espontáneo)	254 (96,6 %)	691 (95,3 %)	0,420	945 (95,6 %)
Hipertensión crónica	9 (3 %)	22 (3 %)	0,918	31 (3 %)
Nulípara	183 (70 %)	399 (55 %)	<0,001	582 (59 %)
Obesidad	47 (18 %)	175 (24 %)	0,046	222 (23 %)
Diabetes pregestacional	4 (2 %)	65 (9 %)	<0,001	69 (7 %)
Años con diabetes	5,8 (3,4)	10,0 (8,2)	<0,001	9,8 (8,1)
Fuma	84 (32 %)	130 (18 %)	<0,001	214 (22 %)
Parto anterior				
Tipo de parto	67 (84 %)	288 (88 %)	0,356	355 (87 %)
Prematuro	19 (27 %)	35 (12 %)	0,003	54 (15 %)
Hipertensión gestacional	9 (14 %)	27 (14 %)	1,0	36 (14 %)
Vol	52,1 (17,4)	66,3 (21,0)	<0,001	62,5 (21,1)
VI	8,6 (5,1)	9,5 (4,4)	<0,001	9,3 (4,6)
FI	47,6 (4,4)	48,8 (4,6)	<0,001	48,5 (4,6)
VFI	4,2 (2,9)	4,7 (2,4)	<0,001	4,6 (2,6)
CRL	61,6 (7,6)	64,8 (7,7)	<0,001	64,0 (7,8)
NT	1,6 (0,4)	1,8 (0,4)	<0,001	1,7 (0,4)
DeltaNT	0,01 (0,4)	0,1 (0,4)	<0,001	0,1 (0,4)
UtPI	2,1 (0,7)	1,8 (0,5)	<0,001	1,9 (0,6)
bhCGMoM	1,2 (0,9)	1,3 (0,9)	<0,001	1,3 (0,9)

Tabla 3.3: Estadística descriptiva en función de si el recién nacido es pequeño o no.

Variables	**PEG n=263**	**no PEG n=725**	**p-valor**	**Total n=988**
PAPPA-MoM	1,0 (0,6)	1,3 (0,7)	<0,001	1,2 (0,7)
Diabetes gestacional	30 (11 %)	159 (22 %)	<0,001	189 (19 %)
Hipertensión gestacional	3 (1 %)	8 (1 %)	1,0	11 (1 %)
Preeclampsia	34 (13 %)	39 (5 %)	<0,001	73 (7 %)
Edad gestacional	273,0 (21,1)	276,3 (13,1)	<0,001	275,4 (15,7)
Tipo de parto	210 (80 %)	600 (83 %)	0,338	810 (82 %)
Peso final	75,4 (14,2)	81,4 (14,7)	<0,001	79,8 (14,8)
Incremento de peso	9,2 (4,9)	10,5 (5,1)	<0,001	10,2 (5,1)
Recién nacido				
Nacido vivo	253 (96 %)	722 (99 %)	<0,001	975 (99 %)
Peso (g.)	2508 (501)	3373,4 (504)	<0,001	3143 (632)
Percentil peso	53,7 (28,4)	3,7 (3,1)	<0,001	40,4 (32,9)
Malformación	4 (2 %)	3 (1 %)	0,160	7 (1 %)
Sexo (masculino)	116 (44 %)	378 (52 %)	0,031	494 (50 %)

PEG: pequeño para la edad gestacional; VI: índice de vascularización; FI: índice de flujo; VFI: índice de vascularización-flujo; CRL: longitud craneal del bebé; NT: longitud de translucencia nucal del bebé; DeltaNT: índice de translucencia nucal; UtPI: índice de pulsatilidad; bhCGMoM: proteína de la gonadotropina coriónica beta-humana; PAPPA-MoM: proteína a-plasmática.

Los resultados del modelo intro son mostrados en la tabla 3.4. Esta tabla proporciona información de los coeficientes de cada variable introducida en el modelo, sus errores estándar, sus estadísticos de Wald, sus p-valores, sus OR y sus intervalos de confianza al 95 % para los OR. Interpretando los p-valores se observa que las variables significativas al 5 % son Estatura, Hipertensión crónica, Nuliparidad, Diabetes pregestacional, Fumadora, Vol, FI, UtPI y PAPPA-MoM. También CRL y DeltaNT presentan un nivel de significación inferior a un 10 %. Los valores de los OR de cada una de las variables y sus respectivos intervalos de confianza al 95 % muestran que son significativas, el 1 no está contenido en el intervalo, de Estatura materna, Hipertensión crónica, Nuliparidad, Fumadora, Vol, FI, UtPI y PAPPA-MoM. La distancia del modelo ajustado (904,65) disminuye con respecto a la del modelo nulo (1144,35) lo que muestra que este modelo aporta información.

La tabla 3.5 muestra la pseudo R^2. Ésta se encuentra entre 0,2 y 0,4 por lo que puede clasificarse como un buen modelo.

Tabla 3.4: Modelo de regresión logística incluyendo todas las variables consideradas (modelo intro).

	Estimación	Error estándar	Z_{Wald}	p-valor	OR	I.C. 95 % OR
Intercepto	14,67	2,873	5,106	<0,001	$2,35\ 10^6$	$[9,14\ 10^3; 7,23\ 10^8]$
Edad materna	-0,003	0,015	-0,187	0,852	0,997	[0,969; 1,027]
Peso materno	-0,016	0,01	-1,593	0,111	0,984	[0,964; 1,004]
Estatura	-0,061	0,016	-3,819	<0,001	0,941	[0,912; 0,971]
Concepción	0,185	0,421	0,439	0,661	1,203	[0,545; 2,89]
Hipertensión crónica	1,108	0,509	2,178	0,029	3,028	[1,08; 8,078]
Nuliparidad	0,603	0,19	3,184	0,002	1,828	[1,264; 2,66]
Obesidad	-0,082	0,347	-0,236	0,813	0,921	[0,465; 1,815]
Diabetes pregestacional	-2,214	0,574	-3,856	<0,001	0,109	[0,03; 0,301]
Fumadora	0,96	0,195	4,93	<0,001	2,61	[1,784; 3,831]
Vol	-0,028	0,005	-5,3	<0,001	0,972	[0,962; 0,982]
VI	-0,09	0,107	-0,837	0,402	0,914	[0,719; 1,079]
FI	-0,059	0,026	-2,246	0,025	0,943	[0,894; 0,991]
VFI	0,154	0,205	0,753	0,452	1,167	[0,845; 1,841]
CRL	-0,022	0,012	-1,794	0,073	0,979	[0,956; 1,002]
DeltaNT	-0,391	0,224	-1,75	0,08	0,676	[0,431; 1,04]
UtPI	0,55	0,153	3,585	<0,001	1,733	[1,285; 2,346]
bhCGMoM	-0,1	0,095	-1,054	0,292	0,905	[0,749; 1,087]
PAPPA-MoM	-0,315	0,15	-2,101	0,036	0,73	[0,54; 0,973]

Tabla 3.5: Pseudo R^2 del modelo intro.

R^2_{MF}	R^2_{CS}	R^2_N
0,2081	0,0005	0,0329

Para ver cuán bueno es este modelo se calcula el estadístico de Hosmer Lemeshow y se observa que ajusta bien los datos ya que 0,469>0,05 (tabla 3.6), es decir, se acepta la hipótesis nula.

Tabla 3.6: Prueba de Hosmer Lemeshow del modelo intro.

χ^2	g.l	p-valor
7,648	8	0,469

Otra forma de ver la bondad de ajuste del modelo consiste en observar una tabla de clasificación (figura 3.1). El 92,4 % de los bebés que no son PEG al nacer fueron clasificados correctamente por el modelo mientras que el 7,6 % no fue clasificado bien, el 36,51 % de los bebés que son PEG fue clasificado correctamente por el modelo frente a un 63,49 % que fue clasificado incorrectamente como no PEG. En global se tiene que el 77,51 % de los recién nacidos están clasificados correctamente.

Tabla 3.7: Tabla de clasificación del modelo intro

Observaciones \ Predicciones	no PEG	PEG	Porcentaje correcto
no PEG	669	55	92,4 %
PEG	167	96	36,51 %

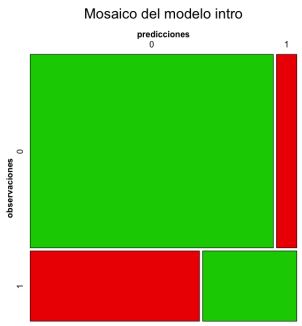


Figura 3.1: Mosaico de la clasificación del modelo intro

Con estos datos se construye la curva ROC (figura 3.2) del modelo, ya que el modelo, con una sensibilidad del 0,3651 (P(clasificado como PEG—cuando es PEG)) y una especificidad del 0,924 (tabla 3.2). Siguiendo el criterio como el área bajo la curva ROC (AUC=0,806) se trata de una discriminación buena debido a que sólo clasifica bien al 35 % de los PEG.

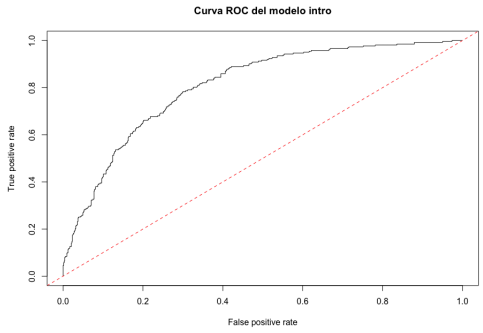
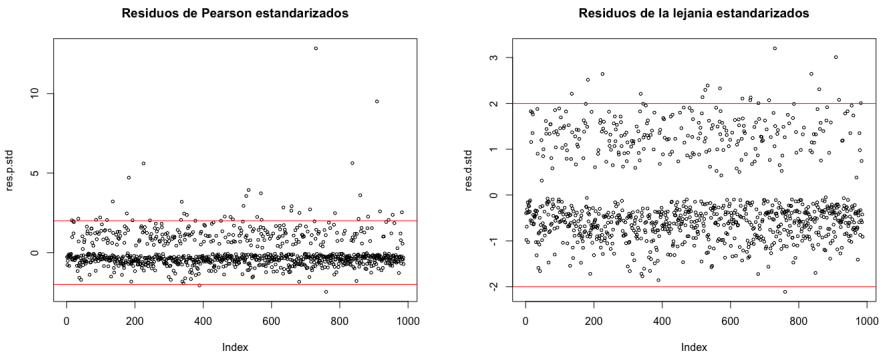


Figura 3.2: Curva ROC del modelo intro

Analizando los residuos de Pearson se obtiene que hay 38 casos cuyo valor absoluto es mayor que 2, que suponen el 3,85 % de los residuos. Y analizando los residuos de la distancia se tiene que 16 casos en valor absoluto son mayores que 2, que son el 1,62 % de los residuos. También se observa en la figura 3.3 que prácticamente no hay ningún recién nacido de peso adecuado clasificado como PEG, lo que implicaría tratar a una madre sin necesidad. Se identifican 5 casos en los que el residuo estandarizado de Pearson es mayor que 4 en valor absoluto. Estos casos son bebés que tuvieron una edad gestacional avanzada que supera las 40

semanas (280 días), excepto un caso que nació por cesárea en la semana 38. Estos bebés fueron clasificados como PEG debido a que su peso fue bajo con respecto a su edad gestacional. En cambio ningún residuo de la lejanía estandarizado supera el 4 en valor absoluto.



(a) Pearson estandarizado (b) Lejania estandarizado

Figura 3.3: Residuos del modelo intro

En cuanto a los outliers, ninguna distancia de Cook ni ningún dfbetas es mayor que 1 en valor absoluto.

A continuación se construye otro modelo utilizando el metodo hacia atrás, es decir, partiendo del modelo anterior se van eliminando las variables menos significativas (Concepción, Obesidad, Edad materna, VFI, VI, bhCGMoM) hasta obtener en la tabla 3.8 el modelo final con las variables significativas (p-in = 0,5 y p-out = 0,1).

Tener peso materno y estatura baja aumentan la probabilidad de tener un recién nacido PEG. Si la madre tiene hipertensión crónica es 2,813 veces más probable que el recién nacido sea PEG que si no la tiene. Si es nulípara es 1,83 veces más probable tener un recién nacido PEG. Si la madre fuma es 2,607 veces más probable que sea PEG que si no fuma. A mayor UtPI es 1,714 veces más probable que el recién sea PEG que si esta disminuye. Por último, la distancia del modelo nulo es 1144,35 mientras que la del modelo ajustado es 908,72. Comparando esta distancia con la del modelo intro se ve que la distancia es un poco más grande lo que indica que sería un poco mejor el modelo intro, pese a que tiene más variables.

El coeficiente pseudo R^2_{MF} se encuentra entre los valores 0,2 y 0,4 (la tabla 3.2), es decir, éste también sería un buen modelo.

La tabla 3.10 de la prueba de Hosmer Lemeshow proporciona un p-valor>0,05 por lo que se acepta la hipótesis nula, es decir, el modelo hacia atrás se ajusta bien a los datos. El 91,85 % de los bebés que no son PEG al nacer fueron clasificados

Tabla 3.8: Modelo de regresión logística eliminando las variables no significativas (modelo hacia atrás).

	Estimación	Error estándar	Z_{Wald}	p-valor	OR	I.C. 95 % OR
Intercepto	14,224	2,623	5,422	<0,001	$1,51\ 10^6$	$[9,35\ 10^3; 2,77\ 10^8]$
Peso Materno	-0,017	0,007	-2,690	0,007	0,983	[0,9703; 0,995]
Estatura	-0,059	0,015	-4,079	<0,001	0,943	[0,916; 0,97]
Hipertensión crónica	1,034	0,504	2,054	0,04	2,813	[1,014; 7,431]
Nulípara	0,605	0,178	3,403	<0,001	1,83	[1,296; 2,602]
Diabetes pregestacional	-2,187	0,572	-3,823	<0,001	0,112	[0,031; 0,308]
Fumadora	0,958	0,19	5,034	<0,001	2,607	[1,796; 3,791]
Vol	-0,029	0,005	-5,395	<0,001	0,972	[0,962; 0,982]
FI	-0,055	0,02	-2,769	0.006	0,947	[0,911; 0,984]
CRL	-0,023	0,012	-1,916	0.055	0,978	[0,955; 1]
DeltaNT	-0,379	0,222	-1,703	0.089	0,685	[0,437; 1,049]
UtPI	0,539	0,151	3,577	<0,001	1,714	[1,278; 2,307]
PAPPAMoM	-0,338	0,147	-2,297	0.022	0,713	[0,53; 0,945]

Tabla 3.9: Pseudo R^2 del modelo hacia atrás

R^2_{MF}	R^2_{CS}	R^2_N
0,2059	0,0005	0,0325

Tabla 3.10: Prueba de Hosmer Lemeshow del modelo hacia atrás.

χ^2	g.l	p-valor
9,7364	8	0,284

correctamente por el nuevo modelo (tabla 3.11), mientras que el 8,15 % no fue clasificado bien, el 35,74 % de los bebés que son PEG fue clasificado correctamente frente a un 64,26 % que fue clasificado incorrectamente como PEG con lo que se repiten los mismos porcentajes del modelo anterior. Puede verse gráficamente en la figura 3.4. En global se tiene que este modelo clasifica un 76,9 % de los recién nacidos correctamente.

Con estos datos se construye la curva ROC (figura 3.5) del modelo hacia atrás con la sensitividad 0,3574 y la especificidad 0,9185. Siguiendo el criterio, como AUC=0,805 y está entre 0,8 y 0,9 se trata de una discriminación buena. Tanto el modelo intro como el modelo hacia atrás discriminan de igual manera.

Analizando los residuos de Pearson se observa que hay 39 casos cuyo valor absoluto es mayor que 2, que suponen el 3,59 % de los residuos. Y analizando los residuos de la distancia se tiene que 19 casos en valor absoluto son mayores que 2, que son el 1,93 % de los residuos. En la figura 3.6a se observan 6 casos en los que el residuo de Pearson estandarizado es mayor que 4 en valor absoluto. Estos casos son bebés que tuvieron una edad gestacional avanzada que supera las 40 semanas (280 días), excepto un caso que nació por cesárea en la semana 38. Estos bebés fueron clasificados como PEG debido a que su peso fue bajo con respecto a su

Tabla 3.11: Tabla de clasificación del modelo hacia atrás

Observaciones \ Predicciones	no PEG	PEG	Porcentaje correcto
no PEG	665	59	91,85 %
PEG	169	94	35,74 %

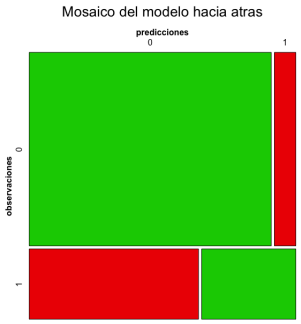


Figura 3.4: Mosaico de la clasificación del modelo hacia atrás

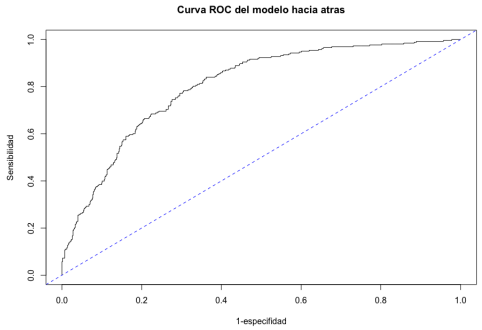
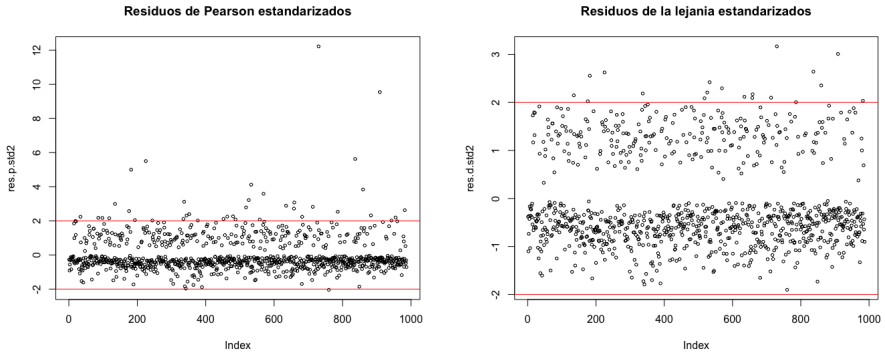


Figura 3.5: Curva ROC del modelo hacia atrás

edad gestacional. En cambio ningún residuo de la lejanía estandarizado supera el 4 en valor absoluto. En cuanto a los outliers, ninguna distancia de Cook ni ningún dfbetas es mayor que 1 en valor absoluto.



(a) Pearson estandarizada (b) Lejania estandarizada

Figura 3.6: Residuos del modelo hacia atrás

3.3. Discriminación de recién nacidos pequeños y grandes para la edad gestacional

En este ejemplo se quiere predecir si el recién nacido va a ser PEG, AEG o GEG utilizando la misma información que en caso anterior. Con la finalidad de tratar a las madres que no vayan a tener recién nacidos con peso adecuado. La variable dependiente Y de este ejemplo tiene 3 categorías:

$$Y = \begin{cases} 0 \text{ si el peso es menor o igual que el percentil 10} \\ 1 \text{ si el peso se encuentra entre el percentil 10 y 90} \\ 2 \text{ si el peso es mayor o igual que el percentil 90} \end{cases} \tag{3.1}$$

Para encontrar un modelo adecuado se seguirá el método hacia atrás, igual que en el ejemplo logístico, partiendo del modelo saturado (3.2).

$$\ln(\frac{\pi_j}{\pi_1}) = \alpha_j + \sum_{k=1}^{17} \beta_{jk} X_k \tag{3.2}$$

Dónde los β_i son los coeficientes del modelo y las X_k las variables incluidos en el modelo. Siendo π_1 la probabilidad de PEG, π_2 la probabilidad de AEG y π_3 la probabilidad de GEG. Se comienza con el modelo de la ecuación anterior y se van eliminando las variables menos significativas, es decir, cuando su p-valor es mayor que 0,1. Realizando este proceso hasta que ya no hay ninguna variable no significativa se llega al siguiente modelo:

$$\ln(\frac{\pi_j}{\pi_1}) = \alpha_j + \sum_{k=1}^{9} \beta_{jk} X_k, \text{ con j=2,3} \tag{3.3}$$

Dónde α_j (intercepto) y β_j (los coeficientes estimados del modelo) se corresponden con las columnas correspondientes de la tabla 3.12.

Tabla 3.12: Modelo final.

Tipo de peso	Variables	Estimación	Error estándar	Z_{Wald}	p-valor	OR	I.C. 95 % OR
	Intercepto	-10,157	1,639	-6,198	<0,001	$3,88\ 10^{-5}$	$[1,6\ 10^{-6}; 9,6\ 10^{-4}]$
	Peso Materno	0,013	0,006	2,053	0,0401	1,013	[1,001; 1,026]
	Estatura	0,066	0,011	5,546	<0,001	1,061	[1,039; 1.084]
	Hipertensión crónica	-1,138	0,498	-2,288	0,0222	0,320	[0,121; 0,85]
AEG	Nuliparidad	-0,517	0,178	-2,915	0,0036	0,596	[0,421; 0,844]
	Diabetes pregestacional	1,839	0,574	3,205	0,0014	6,287	[2,042; 19,356]
	Fumadora	-0,853	0,189	-4,521	<0,001	0,426	[0,294; 0,617]
	Vol	0,034	0,005	6,96	<0,001	1,034	[1,025; 1,044]
	VFI	0,065	0,035	1,842	0,0655	1,067	[0,996; 1,144]
	UtPI	-0,612	0,147	-4,18	<0,001	0,542	[0,407; 0,722]
	Intercepto	-16,111	0,749	-21,503	<0,001	$1,007\ 10^{-7}$	$[2,3\ 10^{-8}$;4,4 10−7]
	Peso Materno	0,036	0,009	3,796	<0,001	1,036	[1,017; 1,056]
	Estatura	0,07	0,008	8,635	<0,001	1,073	[1,056; 1,09]
	Hipertensión crónica	-1,973	0,758	-2,602	0,0093	0,139	[0,032; 0,616]
GEG	Nuliparidad	-1,109	0,273	-4,063	<0,001	0,33	[0,193; 5,632]
	Diabetes pregestacional	3,429	0,624	5,496	<0,001	30,858	[9,083; 104,839]
	Fumadora	-1,312	0,343	-3,829	<0,001	0,269	[0,138; 0,527]
	Vol	0,052	0,007	7,943	<0,001	1,054	[1,04; 1,067]
	VFI	0,103	0,055	1,889	0,0589	1,109	[0,996; 1,235]
	UtPI	-0,884	0,245	-3,604	<0,001	0,413	[0,256; 0,668]

En la tabla 3.12 se observa el resultado del modelo 3.3 con respecto a la categoría PEG. Un coeficiente negativo, o de forma equivalente OR<1, indica que es más probable que se dé esa variable en el grupo de la categoría de referencia (PEG) que en las otras (AEG y GEG). Por otro lado, si es positivo, y por tanto, OR>1, indica que es más probable la variable en cuestión en la categoría (AEG o GEG) que se compara con la de referencia que en esa. Los coeficientes positivos de peso, estatura, diabetes pregestacional, Vol y VFI indican que es más probable en las categorías AEG o GEG que en la PEG. Mientras que los coeficientes de hipertensión crónica, nuliparidad, fumadora y UtPI son negativos lo que quiere decir que es más probable en la categoría PEG que en la AEG y GEG.

La distancia del modelo final es 1478,65 que es menor que la del modelo nulo (1772,45). Para realizar el ajuste global del modelo se utiliza el test de χ^2 de la razón de verosimilitud comparando el modelo final con el saturado, como se obtiene un p-valor=0,984 esto significa que el ajuste del modelo es bueno. También observando los coeficientes pseudo R^2 de la tabla 3.13 no se obtienen valores muy bajos.

Tabla 3.13: Pseudo R^2 del modelo final

R^2_{MF}	R^2_{CS}	R^2_N
0,1658	0,2575	0,3087

Tabla 3.14: Tabla de clasificación del modelo final

Observaciones \ Predicciones	PEG	PAE	GEG	Porcentaje correcto
PEG	94	169	0	35,74 %
PAE	65	535	11	87,56 %
GEG	1	100	12	10,62 %

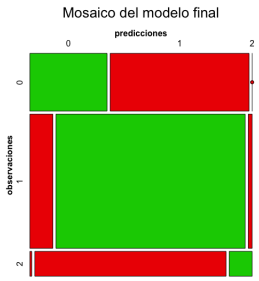


Figura 3.7: Mosaico de la clasificación del modelo final

Observando la tabla 3.14 de observados y predichos se obtiene la tasa de clasificaciones correctas del modelo 64,94 %. Este modelo clasifica correctamente el 35,74 % de los PEG muy parecido al resultado obtenido en el ejemplo anterior, el 87,56 % de los PAE y el 10,62 % de los GEG. El mosaico de la figura 3.7 muestra gráficamente esta información.

No hay ningún residuo mayor que dos en valor absoluto por lo que no se observan datos atípicos.

3.4. Conclusiones

Mediante el modelo de regresión logística se aborda el objetivo de predecir si un recién nacido va a ser PEG o no utilizando información de la madre y obstétricas de los tres primeros meses de embarazo. Para ello, partiendo del modelo intro se construye el modelo hacia atrás. Este modelo consigue un 76,9 % de clasificación

correcta. Aunque la clasificación correcta de PEG es de un 36 %, el cuál no es muy alto, se estaría adelantando la posibilidad de aplicación de tratamiento para corregir esta deficiencia. Además, se observan las siguientes variables como factores de riesgo: hipertensión crónica, nuliparidad, fumadora y UtPI, siendo sus OR con sus intervalos de confianza (I.C. al 95 %) los siguientes: 2,813 con (1,014; 7,431), 1,83 con (1,296; 2,602), 2,607 con (1,796; 3,791) y 1,714 con (1,278; 2,307). Aparecen 5 casos atípicos, esto se debe a la presencia de bebés con edad gestacional superior a las 40 semanas, por lo que estos datos no influyen en el modelo.

Se implementa el modelo multinomial añadiendo una categoría a la variable respuesta, haciendo uso de los mismos datos. Con este modelo se busca predecir si el recién nacido va a ser PEG, AEG o GEG. Este modelo clasifica correctamente un 64,94 %, siendo una tasa de clasificación correcta inferior al modelo de regresión logística. En la categoría GEG, se tiene un 10,62 % de clasificación correcta, pero esto es más complejo utilizando sólo la información de los tres primeros meses, pues uno de los factores que más suelen afectar son cambios hormonales y el desarrollo de diabetes gestacional que suelen aparecer más adelante. Como no hay residuos mayores a uno, no hay existencia de valores atípicos.

A

Anexo

A.1. Código de regresión logística

```
#Regresion logistica ejemplo 1
setwd("~/Desktop/OneDrive - ull.edu.es/TFG/TFG R")
library(readxl)
library(tidyverse)
library(datasets)
library(vcd)
library(ROCR)
library(MASS)
library(ResourceSelection)
library(car)
datos=read_excel("BaseDatosMariano.xlsx")
#Filtramos los datos usando la variable FiltroFinal
datos1=datos %>% filter(FiltroFinal > 0)
attach(datos1)
#Ajuste del modelo logistico intro
Concepcion1=rep(0,length(Concepcion))
for(i in 1:(length(Concepcion))){
  if(Concepcion[i]==1){
    Concepcion1[i]=1
  }
  else{
    Concepcion1[i]=0
  }
}
modelo_glm <- glm(PEG ~ Edad_Materna + Peso_Materno + Estatura +
Concepcion1 + Hipertension_cronica + Nuliparidad + Obesidad +
Diabetes_pregestacional + Fumadora + Vol + VI + FI + VFI + CRL +
```

```
DeltaNT + UtPI + bhCGMoM + PAPPAMoM,data=datos1,
family=binomial(link=logit))
summary(modelo_glm)
#Intervalos del modelo intro
exp(modelo_glm$coefficients)
confint(modelo_glm)
exp(confint(modelo_glm))
#Pseudo R^2 del modelo intro
R_MF_intro=1-(modelo_glm$deviance/modelo_glm$null.deviance)
R_MF_intro
R_CS_intro=(1-(modelo_glm$deviance/modelo_glm$null.deviance)^(2/987))
R_CS_intro
R_N_intro=R_CS_intro/(1-(modelo_glm$null.deviance)^(2/987))
R_N_intro
#Bondad de ajuste del modelo intro
hoslem.test(modelo_glm$model$PEG,fitted(modelo_glm))
#Matriz de confusion del modelo intro
predicciones <- ifelse(test = modelo_glm$fitted.values > 0.5,
yes = 1, no = 0)
matriz_confusion <- table(modelo_glm$model$PEG, predicciones,
                          dnn = c("observaciones", "predicciones"))
matriz_confusion
sum(diag(matriz_confusion))*100/sum(matriz_confusion)
matriz_confusion[1,1]*100/sum(matriz_confusion[1,])
matriz_confusion[2,2]*100/sum(matriz_confusion[2,])
#Mosaico del modelo intro
mosaic(matriz_confusion,main="Mosaico del modelo intro",  shade = T,
 colorize = T,gp = gpar(fill = matrix(c("green3", "red2", "red2",
  "green3"), 2, 2)))
#Curva ROC del modelo intro
suceso=modelo_glm$model$PEG
p=modelo_glm$fitted.values
pred=prediction(p,suceso)
perf=performance(pred,"tpr","fpr")
plot(perf,main="Curva ROC del modelo intro")
abline(0,1,lty=2,col="red")
auc=performance(pred,"auc")
cat("AUC=",auc@y.values[[1]],"\n")
#Residuos del modelo intro
#Residuos de Pearson
```

```
res.p <- residuals(modelo_glm, type = "pearson")
res.p.sig <- abs(res.p) > 2
table(res.p.sig)
res.p.sig <- abs(res.p) > 4
res.orde <- sort(abs(res.p[res.p.sig]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#Residuos de Pearson estandarizados
res.p.std <- rstandard(modelo_glm, type = "pearson")
res.p.std.sig <- abs(res.p.std) > 2
table(res.p.std.sig)
res.p.std.sig <- abs(res.p.std) > 4
table(res.p.std.sig)
res.orde <- sort(abs(res.p.std[res.p.std.sig]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#Residuos de la distancia
res.d <- residuals(modelo_glm, type = "deviance")
res.d.sig <- abs(res.d) > 2
table(res.d.sig)
res.d.sig <- abs(res.d) > 4
table(res.d.sig)
res.orde <- sort(abs(res.d[res.d.sig]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#Residuos de la distancia estandarizados
res.d.std <- rstandard(modelo_glm, type = "deviance") # significativos
res.d.std.sig <- abs(res.d.std) > 2
table(res.d.std)
res.d.std.sig <- abs(res.d.std) > 4
res.orde <- sort(abs(res.d.std[res.d.std.sig]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#gráfico de los residuos del modelo intro
plot(res.p,cex=0.6,main="Residuos de Pearson")
abline(h=c(-2,2),col="red")
```

```
plot(res.p.std,cex=0.6,main="Residuos de Pearson estandarizados")
abline(h=c(-2,2),col="red")
plot(res.d,cex=0.6, main="Residuos de la lejania")
abline(h=c(-2,2),col="red")
plot(res.d.std, cex = 0.6,main="Residuos de la lejania estandarizados")
abline(h = c(-2, 2), col = "red")
#Outliers del modelo intro
medidas.infl_intro <- influence.measures(modelo_glm)
colnames(medidas.infl_intro$infmat)
medidas.infl_intro$infmat
table(medidas.infl_intro$infmat[,22] > 1)
table(medidas.infl_intro$infmat[,23]>(2*p/987))
#Modelo hacia atrás
backwards=step(modelo_glm,direction="backward",stat="wald",alpha=0.1)
summary(backwards)
exp(backwards$coefficients)
confint(backwards)
exp(confint(backwards))
#Pseudo R^2 del modelo hacia atras
R_MF_atras=1-(backwards$deviance/backwards$null.deviance)
R_MF_atras
R_CS_atras=(1-(backwards$deviance/backwards$null.deviance)^(2/987))
R_CS_atras
R_N_atras=R_CS_atras/(1-(backwards$null.deviance)^(2/987))
R_N_atras
#Bondad de ajuste del modelo hacia atras
hoslem.test(backwards$model$PEG,fitted(backwards))
#Matriz de confusion del modelo hacia atras
predicciones1 <- ifelse(test = backwards$fitted.values > 0.5,
 yes = 1, no = 0)
matriz_confusion1 <- table(backwards$model$PEG, predicciones1,
                           dnn = c("observaciones", "predicciones"))
matriz_confusion1
#Porcentaje total
paste("El porcentaje total es:",sum(diag(matriz_confusion1))*100
/sum(matriz_confusion1),"%")
#Porcentaje de verdaderos negativos
paste("El porcentaje de verdaderos negativos es:",matriz_confusion1[1,1]*10
/sum(matriz_confusion1[1,]),"%")
#Porcentaje de verdaderos positivos
```

```
paste("El porcentaje de verdaderos positivos es:",matriz_confusion1[2,2]*100
/sum(matriz_confusion1[2,]),"%")
#Mosaico del modelo hacia atras
mosaic(matriz_confusion1, main="Mosaico del modelo hacia atras", shade = T,colori
 "green3"), 2, 2)))
#Curva ROC del modelo hacia atras
suceso1=backwards$model$PEG
p1=backwards$fitted.values
pred1=prediction(p1,suceso1)
perf1=performance(pred1,"tpr","fpr")
plot(perf1,main="Curva ROC del modelo hacia atras",xlab="1-especifidad",
ylab="Sensibilidad")
abline(0,1,lty=2,col="blue")
auc=performance(pred1,"auc")
cat("AUC=",auc@y.values[[1]],"\n")
#Residuos de Pearson del modelo hacia atras
res.p2 <- rstandard(backwards, type = "pearson")
res.p.sig2 <- abs(res.p2) > 2
table(res.p.sig2)
res.p.sig2 <- abs(res.p2) > 4
table(res.p.sig2)
res.orde <- sort(abs(res.p2[res.p.sig2]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#Residuos de Pearson estandarizados del modelo hacia atras
res.p.std2 <- rstandard(backwards, type = "pearson")
res.p.std.sig2 <- abs(res.p.std2) > 2
table(res.p.std.sig2)
res.p.std.sig2 <- abs(res.p.std2) > 4
table(res.p.std.sig2)
res.orde <- sort(abs(res.p.std2[res.p.std.sig2]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#Residuos de la distancia del modelo hacia atras
res.d2 <- residuals(backwards, type = "deviance")
res.d.sig2 <- abs(res.d2) > 2
table(res.d.sig2)
res.d.sig2 <- abs(res.d2) > 4
```

```
table(res.d.sig2)
res.orde <- sort(abs(res.d2[res.d.sig2]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#Residuos de la distancia estandarizados del modelo hacia atras
res.d.std2 <- rstandard(backwards, type = "deviance") # significativos
table(abs(res.d.std2) > 2)
res.d.std.sig2 <- abs(res.d.std2) > 4
table(res.d.std.sig2)
res.orde <- sort(abs(res.d.std2[res.d.std.sig2]), decreasing = TRUE)
head(res.orde)
head(datos1[names(res.orde), ])
View(head(datos1[names(res.orde), ]))
#gráfico de los residuos del modelo hacia atras
plot(res.p2,cex=0.6,main="Residuos de Pearson")
abline(h=c(-2,2),col="red")
plot(res.p.std2,main="Residuos de Pearson estandarizados",cex=0.6)
abline(h=c(-2,2),col="red")
plot(res.d2,main="Residuos de la lejania",cex=0.6)
abline(h=c(-2,2),col="red")
plot(res.d.std2,main="Residuos de la lejania estandarizados", cex = 0.6)
abline(h = c(-2, 2), col = "red")
#Outliers del modelo hacia atras
medidas.infl2 <- influence.measures(backwards)
colnames(medidas.infl2$infmat)
medidas.infl2$infmat
table(medidas.infl2$infmat[,16] > 1)
table(medidas.infl2$infmat[,17]>(2*p1/987))
```

A.2. Código de regresión multinomial

```
#Regresion multinomial
setwd("~/Desktop/OneDrive - ull.edu.es/TFG/TFG R")
library(readxl)
library(RcmdrMisc)
library(tidyverse)
library(datasets)
library(vcd)
library(nnet)
```

```
datos=read_excel("BaseDatosMariano.xlsx")
#Filtramos los datos usando la variable FiltroFinal
datos1=datos %>% filter(FiltroFinal > 0)
#datos1=filter(datos,FiltroFinal>0) hace lo mismo
attach(datos1)
for(i in 1:(length(PEG))){
  if(Percentil_Peso[i]>=90){
    PEG2[i]=2
  }
   else if(Percentil_Peso[i]<=10){
     PEG2[i]=0
   }
  else
    PEG2[i]=1
}
datos1$PEG2=PEG2
Concepcion1=rep(0,length(Concepcion))
for(i in 1:(length(Concepcion))){
  if(Concepcion[i]==1){
    Concepcion1[i]=1
  }
  else{
    Concepcion1[i]=0
  }
}
#Ajuste de un modelo multinomial
modelo_saturado<-multinom(PEG2 ~ Edad_Materna+Peso_Materno+Estatura+
Concepcion1+Hipertension_cronica+Nuliparidad+Obesidad+
Diabetes_pregestacional+Fumadora+Vol+VI+VFI+CRL+
DeltaNT+UtPI+bhCGMoM+PAPPAMoM,trace=FALSE,
na.action = na.exclude)
summary(modelo_saturado,cor=FALSE,Wald=TRUE)
z = summary (modelo_saturado) $coefficients /
summary (modelo_saturado)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
ic<-confint(modelo_saturado)
exp(ic)
modelo0<-multinom(PEG2~1,data=datos1,trace=FALSE,na.action=na.exclude)
#Tasa de clasificaciones correctas del modelo saturado
```

```
obs<-datos1$PEG2
pre <- predict(modelo_saturado, type="class")
matriz_confusion_modelo_saturado<- table(obs, pre,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo_saturado
tcc=sum(diag(matriz_confusion_modelo_saturado))*100
/sum(matriz_confusion_modelo_saturado)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc,2), "%")
matriz_confusion_modelo_saturado[1,1]*100
/sum(matriz_confusion_modelo_saturado[1,])
matriz_confusion_modelo_saturado[2,2]*100
/sum(matriz_confusion_modelo_saturado[2,])
matriz_confusion_modelo_saturado[3,3]*100
/sum(matriz_confusion_modelo_saturado[3,])
mosaic(matriz_confusion_modelo_saturado,main="Mosaico del modelo saturado",
shade = T,colorize = T,gp = gpar(fill = matrix(c("green3", "red2","red2",
"red2", "green3","red2","red2","red2","green3"), 3, 3)))
#Pseudo R cuadrados del modelo saturado
R_MF_saturado=1-(deviance(modelo_saturado)/deviance(modelo0))
R_MF_saturado
R_CS_saturado=1-exp((deviance(modelo_saturado)-deviance(modelo0))/987)
R_CS_saturado
R_N_saturado=R_CS_saturado/(1-exp(-deviance(modelo0)/987))
R_N_saturado
#Residuos del modelo saturado
residuos=residuals(modelo_saturado)
numSummary(residuos,statistics=c("mean","sd","quantiles"),
quantiles=c(0,0.25,0.5,0.75,1))
#Modelo hacia atras
modelo1=multinom(PEG2 ~ Peso_Materno+Estatura+Concepcion1+
Hipertension_cronica+Nuliparidad+Obesidad+
Diabetes_pregestacional+Fumadora+Vol+VI+VFI+CRL+DeltaNT+
UtPI+bhCGMoM+PAPPAMoM,trace=FALSE,
na.action = na.exclude)
summary(modelo1,cor=FALSE,Wald=TRUE)
z = summary (modelo1) $coefficients /
 summary (modelo1)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
```

```
obs1 <-datos1$PEG2
pre1 <- predict(modelo1, type="class")
matriz_confusion_modelo1<- table(obs1, pre1,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo1
tcc1=sum(diag(matriz_confusion_modelo1))*100
/sum(matriz_confusion_modelo1)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc1,2), "%")
matriz_confusion_modelo1[1,1]*100
/sum(matriz_confusion_modelo1[1,])
matriz_confusion_modelo1[2,2]*100
/sum(matriz_confusion_modelo1[2,])
matriz_confusion_modelo1[3,3]*100
/sum(matriz_confusion_modelo1[3,])
mosaic(matriz_confusion_modelo1,main="Mosaico del modelo 1",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2",
"red2", "red2", "green3","red2","red2","red2","green3"), 3, 3)))
modelo2=multinom(PEG2 ~ Peso_Materno+Estatura+Concepcion1+
Hipertension_cronica+Nuliparidad+Diabetes_pregestacional+
Fumadora+Vol+VI+VFI+CRL+DeltaNT+UtPI+bhCGMoM+
PAPPAMoM,trace=FALSE,na.action = na.exclude)
summary(modelo2,cor=FALSE,Wald=TRUE)
z = summary (modelo2) $coefficients /
summary (modelo2)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
obs2 <-datos1$PEG2
pre2 <- predict(modelo2, type="class")
matriz_confusion_modelo2<- table(obs2, pre2,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo2
tcc2=sum(diag(matriz_confusion_modelo2))*100
/sum(matriz_confusion_modelo2)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc2,2), "%")
matriz_confusion_modelo2[1,1]*100
/sum(matriz_confusion_modelo2[1,])
matriz_confusion_modelo2[2,2]*100
/sum(matriz_confusion_modelo2[2,])
```

```
matriz_confusion_modelo2[3,3]*100
/sum(matriz_confusion_modelo2[3,])
mosaic(matriz_confusion_modelo2,main="Mosaico del modelo 2",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2","red2",
 "red2", "green3","red2","red2","red2","green3"), 3, 3)))
modelo3=multinom(PEG2 ~ Peso_Materno+Estatura+Hipertension_cronica+
Nuliparidad+Diabetes_pregestacional+Fumadora+Vol+VI+VFI+CRL+
DeltaNT+UtPI+bhCGMoM+PAPPAMoM,trace=FALSE,
na.action = na.exclude)
summary(modelo3,cor=FALSE,Wald=TRUE)
z = summary (modelo3) $coefficients /
 summary (modelo3)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
obs3 <-datos1$PEG2
pre3 <- predict(modelo3, type="class")
matriz_confusion_modelo3<- table(obs3, pre3,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo3
tcc3=sum(diag(matriz_confusion_modelo3))*100
/sum(matriz_confusion_modelo3)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc3,2), "%")
matriz_confusion_modelo3[1,1]*100
/sum(matriz_confusion_modelo3[1,])
matriz_confusion_modelo3[2,2]*100
/sum(matriz_confusion_modelo3[2,])
matriz_confusion_modelo3[3,3]*100
/sum(matriz_confusion_modelo3[3,])
mosaic(matriz_confusion_modelo3,main="Mosaico del modelo 3",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2",
"red2","red2", "green3","red2","red2","red2","green3"), 3, 3)))
modelo4=multinom(PEG2 ~ Peso_Materno+Estatura+Hipertension_cronica+
Nuliparidad+Diabetes_pregestacional+Fumadora+Vol+VI+VFI+CRL+
DeltaNT+UtPI+PAPPAMoM,trace=FALSE,na.action = na.exclude)
summary(modelo4,cor=FALSE,Wald=TRUE)
z = summary (modelo4) $coefficients /
 summary (modelo4)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
```

```
obs4 <-datos1$PEG2
pre4 <- predict(modelo4, type="class")
matriz_confusion_modelo4<- table(obs4, pre4,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo4
tcc4=sum(diag(matriz_confusion_modelo4))*100
/sum(matriz_confusion_modelo4)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc4,2), "%")
matriz_confusion_modelo4[1,1]*100
/sum(matriz_confusion_modelo4[1,])
matriz_confusion_modelo4[2,2]*100
/sum(matriz_confusion_modelo4[2,])
matriz_confusion_modelo4[3,3]*100
/sum(matriz_confusion_modelo4[3,])
mosaic(matriz_confusion_modelo4,main="Mosaico del modelo 4",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2",
"red2","red2", "green3","red2","red2","red2","green3"), 3, 3)))
modelo5=multinom(PEG2 ~ Peso_Materno+Estatura+Hipertension_cronica+
Nuliparidad+Diabetes_pregestacional+Fumadora+Vol+VFI+CRL+DeltaNT+
UtPI+PAPPAMoM,trace=FALSE,na.action = na.exclude)
summary(modelo5,cor=FALSE,Wald=TRUE)
z = summary (modelo5) $coefficients /
 summary (modelo5)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
obs5 <-datos1$PEG2
pre5 <- predict(modelo5, type="class")
matriz_confusion_modelo5<- table(obs5, pre5,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo5
tcc5=sum(diag(matriz_confusion_modelo5))*100
/sum(matriz_confusion_modelo5)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc5,2), "%")
matriz_confusion_modelo5[1,1]*100
/sum(matriz_confusion_modelo5[1,])
matriz_confusion_modelo5[2,2]*100
/sum(matriz_confusion_modelo5[2,])
matriz_confusion_modelo5[3,3]*100
```

```
/sum(matriz_confusion_modelo5[3,])
mosaic(matriz_confusion_modelo5,main="Mosaico del modelo 5",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2",
"red2","red2", "green3","red2","red2","red2","green3"), 3, 3)))
modelo6=multinom(PEG2 ~ Peso_Materno+Estatura+Hipertension_cronica+
Nuliparidad+Diabetes_pregestacional+Fumadora+Vol+VFI+CRL+UtPI+
PAPPAMoM,trace=FALSE,na.action = na.exclude)
summary(modelo6,cor=FALSE,Wald=TRUE)
z = summary (modelo6) $coefficients /
 summary (modelo6)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
obs6 <-datos1$PEG2
pre6 <- predict(modelo6, type="class")
matriz_confusion_modelo6<- table(obs6, pre6,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo6
tcc6=sum(diag(matriz_confusion_modelo6))*100
/sum(matriz_confusion_modelo6)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc6,2), "%")
matriz_confusion_modelo6[1,1]*100
/sum(matriz_confusion_modelo6[1,])
matriz_confusion_modelo6[2,2]*100
/sum(matriz_confusion_modelo6[2,])
matriz_confusion_modelo6[3,3]*100
/sum(matriz_confusion_modelo6[3,])
mosaic(matriz_confusion_modelo6,main="Mosaico del modelo 6",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2",
"red2","red2", "green3","red2","red2","red2","green3"), 3, 3)))
modelo7=multinom(PEG2 ~ Peso_Materno+Estatura+Hipertension_cronica+
Nuliparidad+Diabetes_pregestacional+Fumadora+Vol+VFI+UtPI+
PAPPAMoM,trace=FALSE,na.action = na.exclude)
summary(modelo7,cor=FALSE,Wald=TRUE)
z = summary (modelo7) $coefficients /
 summary (modelo7)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
obs7 <-datos1$PEG2
pre7 <- predict(modelo7, type="class")
```

```
matriz_confusion_modelo7<- table(obs7, pre7,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo7
tcc7=sum(diag(matriz_confusion_modelo7))*100
/sum(matriz_confusion_modelo7)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc7,2), "%")
matriz_confusion_modelo7[1,1]*100
/sum(matriz_confusion_modelo7[1,])
matriz_confusion_modelo7[2,2]*100
/sum(matriz_confusion_modelo7[2,])
matriz_confusion_modelo7[3,3]*100
/sum(matriz_confusion_modelo7[3,])
mosaic(matriz_confusion_modelo7,main="Mosaico del modelo 7",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2",
"red2","red2", "green3","red2","red2","red2","green3"), 3, 3)))
modelo8=multinom(PEG2 ~ Peso_Materno+Estatura+Hipertension_cronica+
Nuliparidad+Diabetes_pregestacional+Fumadora+Vol+VFI+UtPI,
trace=FALSE,na.action = na.exclude)
summary(modelo8,cor=FALSE,Wald=TRUE)
z = summary (modelo8) $coefficients /
 summary (modelo8)$ standard.errors
pvalores = (1 - pnorm (abs(z) ,0 ,1))*2
round (pvalores,4)
#modelo final
summary(modelo8,cor=FALSE,Wald=TRUE)
#Intervalos de confianza de los OR del modelo hacia adelante
coeficientes8<-coef(modelo8)
exp(coeficientes8)
ic<-confint(modelo8)
exp(ic)
#Pseudo R cuadrados hacia adelante
R_MF8=1-(deviance(modelo8)/deviance(modelo0))
R_MF8
R_CS8=1-exp((deviance(modelo8)-deviance(modelo0))/987)
R_CS8
R_N8=R_CS8/(1-exp(-deviance(modelo0)/987))
R_N8
#Ajuste global del modelo 8
chi=deviance(modelo8)-deviance(modelo_saturado)
```

```
pchisq(chi,df=8)
#Tasa de clasificaciones correctas del modelo 8
obs8 <-datos1$PEG2
pre8 <- predict(modelo8, type="class")
matriz_confusion_modelo8<- table(obs8, pre8,
dnn = c("observaciones", "predicciones"))
matriz_confusion_modelo8
tcc8=sum(diag(matriz_confusion_modelo8))*100
/sum(matriz_confusion_modelo8)
paste("La tasa de clasificaciones correctas del modelo es:",
round(tcc8,2), "%")
matriz_confusion_modelo8[1,1]*100
/sum(matriz_confusion_modelo8[1,])
matriz_confusion_modelo8[2,2]*100
/sum(matriz_confusion_modelo8[2,])
matriz_confusion_modelo8[3,3]*100
/sum(matriz_confusion_modelo8[3,])
mosaic(matriz_confusion_modelo8,main="Mosaico del modelo final",
shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2",
"red2","red2", "green3","red2","red2","red2","green3"), 3, 3)))
#Residuos del modelo 8
residuos=residuals(modelo8)
numSummary(residuos,statistics = c("mean","sd","quantiles"),
quantiles = c(0,0.25,0.5,0.75,1))
```

Bibliografía

[1] Silva Ayçaguer, L.C. y Barroso, I.M.(2004) *Regresión logística (Cuadernos de Estadística)*. Reading: La Muralla.

[2] Agresti, A.(2007) *An introduction to Categorical Data Analysis*. Reading: John Wiley & Sons.

[3] Green, DM. Swets, JA. (1966) *Signal Theory and psichophysics*. Reading: John Wiley & Sons.

[4] McCullagh, P. y Nelder, J.(1989). *Generalized Linear Models*. Recuperado de: http://www.utstat.toronto.edu/~brunner/oldclass/2201s11/readings/glmbook.pdf.

[5] Hosmer, David W. y Lemeshow , S.(2000) *Applied Logistic Regression*. Recuperado de: http://resource.heartonline.cn/20150528/1_3kOQSTg.pdf.

[6] Del Valle Benavides, A.R. (2017). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones*(Trabajo final de grado). Universidad de Sevilla, Sevilla, España.

[7] Dueñas Rodríguez M.A.(2011). *Modelos de respuesta discreta en R y aplicación con datos reales*(Trabajo final de master). Universidad de Granada, Granada, España.

[8] Pando, V. y San Martín, R. (2004). *Regresión logística multinomial*. Cuadernos de la Sociedad Española de Ciencias Forestales. 18, 323-327.

[9] Gómez-Roig, M. D. (2012). *Pequeño para la edad gestacional (PEG) desde el período prenatal hasta la adolescencia*. Revista Española de Endocrinología Pediatrica, 3(2), 87–89, doi:10.3266.

[10] Unceta-Barrenechea, A., Aguirre Conde, A., Pérez Legórburu, A., Echáriz Urcelay, I. (2008). *Recién nacido de peso elevado*. Protocolos Diagnóstico Terapéuticos de la AEP: Neonatología. 10, 85–90. Recuperado de https://www.aeped.es/sites/default/files/documentos/10_1.pdf.

[11] R Core Team. *R: A language and environment for statistical computing*. 2019. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Cathegorical data analysis: logistic and multinomial regression

Sección de Matemáticas
Universidad de La Laguna

Mariano Brage Escalona

Facultad de Ciencias · Sección de Matemáticas
Universidad de La Laguna
emailAuTFG@ull.edu.es

Abstract

Currently, the study of categorical dependent variables is useful in dif- ferent fields such as health, psychology, sociology and so on. Two re- gression models on categorical variable answers will be developed in this essay, the first being the logistic regression model, which has one dichotomous variable, whilst the second is the multinomial model, in which the variable is polytomous. These two models will be implemen- ted using real gynaecological data with the intention of generating pre- dictions related to the weight of the babies from the maternal informa- tion and obstetrical variables gathered in the first medical examination in the first trimester of the pregnancy.

Introduction

The logistic regression and multinomial regression models are techniques that allow to relate a categorical variable with a set of independent variables to predict events. From the coefficients of the model, it can be interpreted the effects these variables have on the response. The Generalized Linear Model (GLM) unifies the regression models linear, logistic and Poisson. This helps the response variables have a different distribution of errors than normal. The logistic regression is within the GLM, using as link function the logit, and this is considered an extension of linear regression models, with the variable dichotomous response, with the particularity that the path of the function is bounded in the interval [0,1] and, on the other hand, the error estimation procedure is maximum-verisimilitude. The multinomial regression model is a generalization of the previous one in which the response variable is polytomic.

1. Logistic regression

This chapter introduces the logistic regression model. This is used when the response variable, Y, is dichotomous. It is thus formulated:

$$\ln(\frac{\pi}{1-\pi}) = \alpha + \sum_{k=1}^{K} \beta_k x_k \tag{1}$$

The likehood ratio (RV) or the Wald statistician (Z_{Wald}) allows us to check if a variable is significant for the model.

$$RV = -2\ln(\frac{V_0}{V}) \tag{2}$$

$$Z_{Wald} = \frac{b}{se(b)} \tag{3}$$

To check the efficiency of the model, there are different measures of adjustment such as the Hosmer-Lemeshow statistic, the pseudo R2, the classification table and the ROC curve, specifically the area under the ROC curve (AUC) that is better the closer to 1 it is.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} + \sum_{i=1}^{k} \frac{(O_i^* - E_i^*)^2}{E_i^*} \tag{4}$$

Table 1: *Clasification table*

	$Z=1$ (Sick)	$Z=0$ (healthy)
$Y=1$ (test+)	V_+	F_+
$Y=0$ (test-)	F_-	V_-

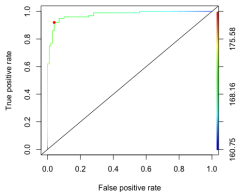


Figure 1: *ROC curve*

To identify possible outliers, an analysis of the residuals is made using the Pearson residuals, Pearson standardized, the distance, the standardized distance, the Cook distance, the influence value and the dfbetas.

2. Multinomial regression

The second chapter generalizes the regression model logistics by adding one or more categories to Y. It comes from the following equation.

$$\ln(\frac{\pi_j}{\pi_J}) = \alpha_j + \sum_{k=1}^{K} \beta_{jk} X_{jk} \tag{5}$$

It is checked if a variable is significant for the model using Wald's statistic and the conditional contrast of the reason of verisimilitude. The efficiency of the model is analyzed using the likehood ratio and the correct classification rate. The latter consists of comparing the results obtained with those observed.

3. Applications of logistic and multinomial regression models

In the final chapter, these two models are implemented to a sample of data affected by the characteristics of pregnant women. Through the logistic regression model, it is inferred based on the data from the first three months, whether the newborn will be SGA (Small for Gestational Age) or not SGA. Employing the multinomial model, a prediction is made, based on the previous data, about whether the newborn is SGA, AWE (Adequate weight for gestational age) and LGE (Large for gestational age).

References

[1] Silva Ayçaguer, L.C. y Barroso, I.M.(2004) *Regresión logística (Cuadernos de Estadística)*. Reading: La Muralla.

[2] Agresti, A.(2007) *An introduction to Categorical Data Analysis*. Reading: John Wiley & Sons.

[3] Green, DM. Swets, JA. (1966) *Signal Theory and psichophysics*. Reading: John Wiley & Sons.

[4] McCullagh, P. y Nelder, J.(1989). *Generalized Linear Models*. Recuperado de: *http://www.utstat.toronto.edu/ brunner/oldclass/2201s11/readings/glmbook.pdf*.

[5] Hosmer, David W. y Lemeshow , S.(2000) *Applied Logistic Regression*. Recuperado de: *http://resource.heartonline.cn/20150528/1_3kOQSTg.pdf*.