Regresión Logística

Autor: Lic. Verónica Cuenca Ramallo

Un método conocido en un modelo de regresión es la regresión logística, este método esta destinado a utilizarlo cuando se presente datos binomiales (dicotómicas) o multinomiales (multivariadas, politomicas).

En la actualidad existe muchos estudios sobre lo que es la regresión logística y sus aplicaciones en diferente áreas, en este articulo se encargara esta regresión.

1. Introducción

En estadística, la regresión logística es un modelo de regresión para variables dependientes, es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. Es un modelo lineal generalizado que usa como función de enlace la función logit.

La regresión logística es usada extensamente en las ciencias médicas, sociales y otros, esto significa áreas del conocimiento humano donde las variables que se analizan son en su mayoría cualitativas.

Otros nombres para la regresión logística usados en varias áreas de aplicación incluyen:

1.1 Modelo logístico, modelo logit, y clasificador de máxima entropía.

En los modelos de regresión se quiere conocer la relación entre:

- Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial).
- Una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas.

Siendo la ecuación inicial del modelo de tipo exponencial, si bien su transformación logarítmica (logit) permite su uso como una función lineal.

Como vemos, las covariables pueden ser cuantitativas o cualitativas. Las covariables cualitativas deben ser dicotómicas, tomando valores 0 para su ausencia y 1 para su presencia (esta codificación es importante, ya que cualquier otra codificación provocaría modificaciones en la interpretación del modelo).

Pero si la covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo deberíamos realizar una transformación de la misma en varias covariables cualitativas dicotómicas ficticias o de diseño (las llamadas variables dummy), de forma que una de las categorías se tomaría como categoría de referencia. Con ello cada categoría entraría en el modelo de forma individual.

En general, si la covariable cualitativa posee n categorías, habrá que realizar n-1 covariables ficticias.

En resumen el modelo de regresión logística se utiliza cuando estamos interesados en pronosticar la probabilidad de que ocurra o no un suceso determinado.

Por ejemplo, a la vista de un conjunto de pruebas médicas, que una persona tenga una determinada enfermedad, o bien que un cliente devuelva un crédito bancario, el comportamiento electoral, color de ojos etc.

2. El modelo de la regresión logística

El modelo de regresión en el caso multivariante se considera una agrupación de p variables independientes cual esta denotado a voluntad por el vector $x`=(x_1,x_2,\ldots\ldots..,x_k)$.
Se asume que cada uno de estas variables es una escala mínima de intervalo, permite la probabilidad condicional que la respuesta se presenta con la notación siguiente:

$$P(Y=1|\ x)=\pi(x)$$

El modelo de regresión logística (logit) múltiple esta dada por la siguiente ecuación.

$$Y= \beta_0+\beta_1 x_1 +\beta_2 x_2 +\beta_3 x_3 +\ldots\ldots\ldots\ldots+\beta_k x_k +\mu$$

Donde:

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

La regresión logística analiza datos distribuidos binomialmente de la forma:

$$Y \sim B(p_i, n_i), i = 1,2,\ldots,m$$

$$Y \sim B(p_i, n_i), i = 1,2,\ldots\ldots\ldots,m$$

Donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas.

El modelo es entonces obtenido en base a lo que cada ensayo (valor de i) y el conjunto de variables explicativas/independientes puedan informar acerca de la probabilidad final. Estas variables explicativas pueden pensarse como un vector X_i k-dimensional y el modelo toma entonces la forma

$$p_i = E(\frac{Y_i}{n_i}|X_i)$$

Los logits de las probabilidades binomiales desconocidas es decir los logaritmos de los odds o ratios de probabilidad que indica cuanto se modifican las probabilidades por unidad de cambio en las variables x, son modeladas como una función lineal de los x_i.

$$logit\ (p_i) = Ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i}$$

Es importante notar que el elemento particular x_i puede ser ajustado a 1 para todo i obteniéndose un intercepto en el modelo. Los parámetros desconocidos β_j son usualmente estimados a través de máximo verosimilitud.

La interpretación de los estimados del parámetro β_j es como los efectos aditivos en el log odds ratio para una unidad de cambio en la j-ésima variable explicativa.

En el caso de una variable explicativa dicotómica, por ejemplo género, e^{β} es la estimación del odds ratio de tener el resultado para, poder decir algo, hombres comparados con mujeres.

El modelo tiene una formulación equivalente dada por:

$$p_i = \frac{1}{1 + e^{-\beta_0+\beta_1 x_{1,i}+\cdots..+\beta_k x_{k,i})}}$$

Esta forma funcional es comúnmente identificada como una red neuronal de una sola capa, calcula una salida continua en lugar de una función por partes. La derivada de p_i con respecto a $x = x_1 \ldots x_k$ es calculada de la forma general:

$$y = \frac{1}{1+e-f(X)}$$

Donde $f(X)$ es una función analítica en X.

3. Algunos aspectos a tener en cuenta para el uso de la regresión logística.

- Tamaño de la muestra y número de variables independientes.
- Una de las ventajas de la regresión logística es que permite el uso de múltiples variables con pocos casos esto es relativo, sin embargo hay que tener en cuenta algunas precauciones.
- Sugerencia el número de sujetos debe se superior a 10(k+1) donde K es el número de variables explicativas.

- Si se introducen variables Dummy, el número de elementos en la muestra debe aumentar.

- Si una de las variables es dicotómica, además de respuesta esta no tiene al menos 10 casos en cada uno de sus dos respuestas posibles ocurre que sus estimaciones no son confiables.

- Cuando las variables independientes se presenta en gran número, puede indicar que no se ha reflexionado suficientemente sobre el problema.

- Se debe tener en cuenta el efecto sobre el riesgo de que ocurra el evento, de los cambios de las variables explicativas cuando son cuantitativas (continuas).

- Cuando algunas de las variables independientes analizadas están altamente correlacionadas, los resultados que se obtienen pueden no ser satisfactorios.

- Para no tener problemas en la aplicación primero se debe realizar un análisis previo univariado entre las distintas variables explicativas.

- Para que la regresión logística tenga un sentido claro, debe existir una relación monótona entre las variables explicativas y la de respuesta, esto significa que el aumento de las unas se acompañe del aumento a lo disminución aproximadamente constante de la otra, para todo el rango de valores estudiados.

4. Algunas observaciones de tipo práctico

- Para una mejor interpretación de los coeficientes cualesquiera β_1 es necesario referirnos al concepto riesgo relativo.

- El exponencial de los β_1 se corresponde con el riesgo relativo, o sea, es una medida de la influencia de la variable x_i sobre el riesgo de que ocurra ese hecho y suponiendo que el resto de las variables del modelo permanezcan constantes.

- Una vez estimados los valores de β_0 y β_1, podemos determinar la probabilidad del suceso para distintos valores de los X_i.

- Para las variables explicativas (categóricas) ya sean nominales u ordinales de más de 2 categorías (politómicas), para incluirlas en el modelo hay que darles un tratamiento especial.

- Si estamos en presencia de una variable nominal con C categorías, debemos incluirla en el modelo de regresión logística como variable categórica, de manera que a partir de ella se crean C-1 variables dicotómicas llamadas *dummy* se debe precisar con cuál de las categorías de la variable original interesa comparar el resto y esa será la llamada categoría de referencia.

- En el caso de las variables ordinales se puede asumir que la escala funciona aproximadamente a un nivel cuantitativo, se pueden manejarse como variables *dummy*.

5. Bibliografía.

[1] David W.Hosmer (1989) "APPLIED LOGISTIC REGRESSION "

[2] David W.Hosmer (2000) "LOGISTIC REGRESSION"

NOTA.-El ejemplo aplicativo se realizará en la siguiente revista Nº 9 se tratará para casos dicotómicos y casos multivariados, también se explicará la terminología.