Modelos de elección discreta logit (Modelo de Regresión Logística)

Introducción

Si $Y_i \sim \text{Binomial}(1, P_i) \rightarrow f(y) = P_i^{y_i}(1-P_i)^{1-y_i}$, $y_i = 0,1$ se dice que Y_i tiene distribución binomial puntual o de Bernoulli. Se denomina odds (o ventaja) al cociente $\pi_i = \frac{P_i}{1-P_i}$ e indica cuánto más o menos probable es el éxito que el fracaso. En consecuencia como

$$\pi_i = \frac{P_i}{1-P_i} \rightarrow P_i = \frac{\pi_i}{\pi_i + 1} = \frac{\text{odds}_i}{\text{odds}_i + 1}$$

Ejemplo

Si la probabilidad de que un joven padezca de ludopatía es $P_i = 0.8$ entonces el odds_i será $\pi_i = \frac{P_i}{1-P_i} = \frac{0.8}{0.2} = 4$ y se interpreta diciendo que el éxito (tener ludopatía) es 4 veces más probable que el fracaso (no padecer ludopatía).

Como $\pi_i = 4 \rightarrow P_i = \frac{\pi_i}{1+\pi_i} = \frac{4}{4+1} = 0.8$.

Observaciones

Se denomina odds ratio (o riesgo relativo o ventaja comparativa) de dos sucesos a $\theta_{1,2} = \frac{\pi_1}{\pi_2}$. Si $\theta_{1,2} = \frac{\pi_1}{\pi_2} > 1$, el éxito es más ventajoso que el fracaso en el primer suceso. Por ejemplo si el odds de un varón respecto a la ludopatía es $\pi_1 = 4$, y el odds para una joven es $\pi_1 = 1.6$ entonces el odds ratio del varón respecto a la chica será $\theta_{1,2} = \frac{\pi_1}{\pi_2} = 2.5 > 1$, y se interpretará de la siguiente manera: padecer ludopatía es 2.5 más ventajoso ("probable") en varones. Si $\theta_{1,2} = \frac{\pi_1}{\pi_2} < 1$, el éxito es menos ventajoso que el fracaso en el segundo suceso. Si $\theta_{1,2} = \frac{\pi_1}{\pi_2} = 1$, los odds en ambos sucesos serán iguales.

Modelos de Regresión Logística

Un modelo de regresión con variable dependiente binomial (o multinomial) será un modelo que permita estudiar si dicha variable discreta depende o no, de otra u otras variables.

1. Los Modelos Dicotómicos

Se considera que la variable dependiente Y o clase es una variable dicotómica que toma dos alternativas, de tal manera que cada individuo de la muestra tiene que pertenecer a una y sólo una, de estas alternativas (clases o grupos).

Es posible representar a una variable dicotómica de la siguiente manera:

$$Y_i = \begin{cases} 1 & \text{Prob}\left(Y_i = 1\right) = P_i \\ 0 & \text{Prob}\left(Y_i = 0\right) = 1 - P_i \end{cases} \qquad (1)$$

De lo anterior se puede hallar el valor esperado de Y_i de la siguiente manera:

$$E\left(Y_i\right) = 1 \times P_i + 0 \times \left(1 - P_i\right) = P_i \qquad (2)$$

Asumiendo que Y_i es explicado por un conjunto de regresores (o variables independientes que pueden ser cuantitativas o cualitativas) $X_{2i}, X_{3i}, \ldots, X_{ki}$. Se le llama Z_i a la siguiente función lineal de estos regresores:

$$Z_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} = \begin{bmatrix} 1 & X_{2i} & \cdots & X_{ki} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = x_i' \beta \quad (3)$$

En la igualdad anterior los β_j son parámetros desconocidos.

Por lo tanto, se puede representar el valor esperado de Y_i condicionada a las variables explicativas (o regresores) de la siguiente manera.

$$E\left(Y_i / X_{2i}, X_{3i}, \cdots, X_{ki}\right) = F\left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}\right) = F\left(Z_i\right) \quad (4)$$

En consecuencia el Modelo Estocástico (o Aleatorio) puede ser representado así:

$$Y_i = E\left(Y_i / X_{2i}, X_{3i}, \cdots, X_{ki}\right) + u_i = F\left(Z_i\right) + u_i \qquad (5)$$

En la igualdad anterior u_i es una perturbación aleatoria (o estocástica).

Dependiendo cuál sea la función F seleccionada se obtienen diferentes Modelos Dicotómicos.

Se considerará el Modelo Logit (o sea cuando F es la Distribución Logística).
Con la finalidad de que las predicciones del Modelo Dicotómico tengan relación lógica con los aspectos teóricos se tiene que establecer las siguientes propiedades:

$$\lim_{Z_i \to +\infty} \text{Prob}\left(Y_i = 1\right) = 1$$
$$\lim_{Z_i \to -\infty} \text{Prob}\left(Y_i = 0\right) = 0$$

Lo anterior se cumple perfectamente si F es una función de distribución aleatoria. Por lo tanto, si F es la Distribución Logística se obtiene el Modelo Logit pero si F es la Distribución Normal estándar se obtiene el Modelo Probit. Es posible considerar otras distribuciones, pero generalmente se usan las dos mencionadas anteriormente. Se utilizará la Distribución Logística (Modelo Logit).

En un Modelo Logit se cumple lo siguiente:

$$P_i = E\left(Y_i / X_{2i}, X_{3i}, \cdots, X_{ki}\right) = F\left(Z_i\right) = \frac{1}{1+e^{-Z_i}} = \frac{e^{Z_i}}{1+e^{Z_i}} = \Lambda\left(Z_i\right) \quad (6)$$

En la igualdad anterior $\Lambda(.)$ es la notación usual para la función de Distribución Acumulativa Logística.

La Función de Densidad Logística está dada por:

$$f\left(Z_i\right) = \Lambda\left(Z_i\right)\left(1-\Lambda\left(Z_i\right)\right) \qquad (7)$$

Con los Modelos de Elección Discreta es común el uso del concepto de razón de apuestas (en inglés odd) en lugar que el concepto de probabilidad. La razón de apuestas (odd) es una relación entre dos probabilidades. Es la razón entre la probabilidad de que se produzca un suceso y la probabilidad de que no se produzca ese suceso. La expresión de razón de apuestas tiene el mismo significado que en el juego, es decir, si en una apuesta donde existen solamente dos posibilidades, se dice que las apuestas están a la par, eso quiere decir que la razón de apuestas será 1:1. Hay que tener en cuenta que la razón de apuestas no es nunca una probabilidad, sino una razón de probabilidades, con base a lo anterior se hará su deducción para el Modelo Logit. La probabilidad de que el suceso no tome el valor 1 es:

$$1-P_i = 1-\frac{1}{1+e^{-Z_i}} = \frac{1+e^{-Z_i}-1}{1+e^{-Z_i}} = \frac{1}{e^{Z_i}\left(1+e^{-Z_i}\right)} = \frac{1}{1+e^{Z_i}}$$

En consecuencia, la razón de apuestas (odd), se da por la siguiente expresión:

$$\pi_i = \frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = \frac{1+e^{Z_i}}{1+\frac{1}{e^{Z_i}}} = e^{Z_i} = e^{\beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+\ldots+\beta_k X_{ki}} \qquad (8)$$

NOTA: Asumiendo que sólo hay una variable independiente X_2 que sólo puede tomar valores 0 o 1, la ventaja u odds de la opción 1 $\left(Y_i = 1\right)$ para $x_2 = 1$ será $\pi_1 = \frac{P\left(Y_i = 1\right)}{1-P\left(Y_i = 1\right)} = \frac{P_1}{1-P_1}$, y para $x_2 = 0$, $\pi_0 = \frac{P\left(Y_i = 0\right)}{1-P\left(Y_i = 0\right)} = \frac{P_0}{1-P_0}$. Entonces el odds ratio (o riesgo relativo o ventaja comparativa) de $x_2 = 1$ respecto a $x_2 = 0$ será:

$$\theta_{1,0} = \frac{\pi_1}{\pi_0} = \frac{\frac{e^{\beta_1+\beta_2}/\left(1+e^{\beta_1+\beta_2}\right)}{1/\left(1+e^{\beta_1+\beta_2}\right)}}{\frac{e^{\beta 1}/\left(1+e^{\beta_1}\right)}{1/\left(1+e^{\beta_1}\right)}} = \frac{e^{\beta_1+\beta_2}}{e^{\beta_1}} = e^{\beta_2}$$

Este resultado nos indica cómo la razón de apuestas de observar $Y_i = 1$ cambia ante un incremento unitario en la variable X_2. Cuando $e^{\beta_2} > 1$ la variable X_2 incrementa la razón de apuestas de observar $Y_i = 1$. Cuando $e^{\beta_2} < 1$ la variable X_2 reduce la razón de apuestas de observar $Y_i = 1$.

VOLVIENDO AL MODELO GENERAL

Con esta clase de modelos, en lugar de la razón de apuestas (odds), se usa el logaritmo neperiano de esa razón. Por lo tanto y teniendo en cuenta (3):

$$\ln \pi_i = \ln\left[\frac{P_i}{1-P_i}\right] = Z_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_j \qquad (9)$$

Se aprecia que, por una parte la razón de apuestas es una función no lineal de los parámetros β, de otro lado el logaritmo de la razón de apuestas si es una función lineal de los β.

Derivando: $\partial \frac{\ln \pi_i}{\partial X_h} = \beta_h$ entonces β_h es aproximadamente el incremento del logaritmo de la razón de apuestas cuando la variable X_h se incrementa en una unidad permaneciendo constantes las demás variables.

Supuestos

Las perturbaciones μ_i son homoscedásticas y no autocorrelacionadas.

Estimación de los parámetros β_j

El modelo logit es no lineal. El método de estimación es el de máxima verosimilitud y con procedimientos iterativos.

Contraste de significación para un β_j

Para probar $\begin{matrix} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{matrix}$ se utiliza el estadístico de Wald.

Contraste de significación para todos los β_j

Para verificar: $\begin{matrix} H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \\ H_1 : \text{Al menos un } \beta_j \neq 0 \end{matrix}$ se utiliza el estadístico de razón de verosimilitud $RV_0 = -2\left[\ln L_0 - \ln L\right]$, donde $\ln L$ es el logaritmo de la función de verosimilitud que se ha obtenido al estimar el modelo completo, mientras que $\ln L_0$ es el logaritmo de la función de verosimilitud al estimar el modelo con sólo el término independiente. Se cumple que $RV_0 = \ln L_0 - \ln L \sim \chi^2_{(k-1)}$ cuando el tamaño de muestra (n) tiende al infinito.

Prueba de bondad de ajuste de la Devianza

H_0 : El modelo se ajusta bien a los datos . El estadístico de prueba es:

$$D = -2\left[\frac{\text{F. de verosimilitud del modelo seleccionado}}{\text{F. de verosimilitud del modelo saturado}}\right] \sim \chi^2_{(n-k)}$$

Prueba de bondad de ajuste de Hosmer Lemeshow

H_0 : El modelo ajustado es el adecuado o que no existen diferencias entre valores observados y predichos.

Para esta prueba se dividen todos los casos en deciles basados en probabilidades predichas. A partir de las probabilidades predichas y de los datos observados se considera el siguiente estadístico: $HL = \sum_{i=1}^{10} \frac{\left[O_i - N_i \overline{\pi}_i\right]^2}{N_i \overline{\pi}_i \left(1 - \overline{\pi}_i\right)} \sim \chi^2_8$ donde:

O_i : es el número de unos del decil i.

$\overline{\pi}_i$: es la media de las probabilidades predichas en el decil i.

N_i : es el número de observaciones del decil i.

2. Los Modelos Logit Multinomiales

En los casos anteriores se han considerado Modelos de Elección Discreta en los que existen dos alternativas (clases o grupos). Pero muchas veces son más de dos las alternativas de donde elegir para tomar una decisión. Para resolver este problema aparecen los Modelos Multinomiales.

A los Modelos de Elección Discreta con más de dos alternativas (clases o grupos) se les llaman Modelos Multinomiales. Con los Modelos Multinomiales ocurre lo mismo que con los Modelos Dicotómicos, se puede utilizar Modelos Logit y Modelos Probit , aunque los Modelos Logit (asumiendo Distribución Logística) son los que más se usan, tanto en la literatura así como en los programas estadísticos.

En el Modelo Logit se considera que el número de alternativas o clases es J+1 (0, 1, 2, ..., J), pero considerando la alternativa 0 como la categoría de referencia. Como se apreció, en los Modelos Dicotómicos sólo hay un vector de parámetros. Entonces, para trabajar con el Modelo Logit Multinomial son necesarios J vectores de parámetros. El número de vectores de parámetros es igual al número de categorías sin considerarse la categoría de referencia.

Supóngase que se tiene la siguiente relación,

$$Z_{ij} = \beta_{1j} + \beta_{2j} X_{2i} + \beta_{3j} X_{3i} + \ldots + \beta_{kj} X_{ki} = \begin{bmatrix} 1 & X_{2i} & \cdots & X_{ki} \end{bmatrix} \begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \vdots \\ \beta_{kj} \end{bmatrix} = x_i' \beta_j$$

Entonces las probabilidades de cada alternativa o clase en un Modelo Logit Multinomial son representadas como sigue:

$$P_{ij} = \text{Prob}\left(Y_i = j\right) = \frac{e^{x_i' \beta_j}}{1 + \sum_{K=1}^{J} e^{x_i' \beta_K}} \quad j = 1, 2, \cdots, J$$

$$P_{i0} = \text{Prob}\left(Y_i = 0\right) = \frac{1}{1 + \sum_{K=1}^{J} e^{x_i' \beta_K}} \qquad (10)$$

Hay que notar que si J es igual a uno, el Modelo Multinomial es igual al Dicotómico.
El logaritmo neperiano de la razón de apuesta entre la alternativa j y la alternativa de la categoría de referencia (0) está dada por:

$$\ln\left(\frac{P_{ij}}{P_{i0}}\right) = x_i'\beta_j = Z_{ij} = \beta_{1j} + \beta_{2j}X_{2i} + \beta_{3j}X_{3i} + \ldots + \beta_{kj}X_{ki} \qquad (11)$$

Es posible hallar el logaritmo de la razón de apuestas entre cualquier par de alternativas o clases. Como, por ejemplo, entre las alternativas j y h se tendrá:

$$\ln\left(\frac{P_{ij}}{P_{ih}}\right) = x_i'\left(\beta_j - \beta_h\right) \qquad (12)$$

Ejemplo de Regresión Logística Dicotómica (Archivo: Empresas1)

En un estudio de mercado se desea investigar los principales factores que pueden influir en aumentar la probabilidad de que un nuevo producto sea introducido con éxito en el mercado. Con esta finalidad, se ha aplicado una encuesta a 240 empresas industriales de las cuales 156 declararon haber intentado introducir en el mercado un nuevo producto. Entonces, el objetivo es explicar el comportamiento en términos de probabilidad de una variable dependiente dicotómica (éxito o fracaso en el lanzamiento de un nuevo producto), en función de un conjunto de variables predictoras.

Utilizando el Método Introducir y Categóricas: Último se obtiene:

Prueba Ómnibus

$H_0 : \beta_2 = \beta_3 = \ldots = \beta_7 = 0$

$H_1 :$ Al menos un $\beta_j \neq 0$

Valor calculado 107.482 y pvalor=0

Prueba de Hosmer Lemeshow

$H_0 :$ El modelo ajustado es el adecuado o que no existen diferencias entre valores observados y predichos.

pvalor=

Pruebas de Wald

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

De abajo para arriba:

pvalores= 0, 0.005, 0.426, 0.009, 0.01, 0.009, 0.834, 0.022, 0

Algunas interpretaciones

$\hat{\beta}_1 = -6.902$, es el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ cuando las otras variables toman el valor de cero.

$\hat{\beta}_2 = 2.021$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ se incrementa en 2.021cuando el gasto en publicidad aumenta en 1 um. y las demás variables permanecen constantes.

$\hat{\beta}_3 = 1.231$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ se incrementa en 1.231cuando el grado de novedad es por mejoras sustanciales y las demás variables permanecen constantes.

$\hat{\beta}_6 = -2.156$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ decrece en 2.156 cuando Intensidad Tecnológica del Sector de Actividad de la Empresa es baja y las demás variables permanecen constantes.

$\hat{\beta}_7 = -0.687$, el valor del $\ln\left(\frac{\hat{P}_1}{1-\hat{P}_1}\right)$ decrece en 0.687 cuando Intensidad Tecnológica del Sector de Actividad de la Empresa es media y las demás variables permanecen constantes.

$e^{\hat{\beta}_2} = 7.543$, el Resultado de introducir un nuevo producto al mercado como éxito es 7.543 más ventajoso (probable) por cada um de gasto en publicidad.

$e^{\hat{\beta}_4} = 1.177$, el Resultado de introducir un nuevo producto al mercado como éxito, para Tipo del producto de consumo industrial es 1.177 más ventajoso (probable) que para consumo final.

$e^{\hat{\beta}_7} = 0.503$, el Resultado de introducir un nuevo producto al mercado como éxito, cuando la intensidad tecnológica del sector de actividad de la empresa es media es 0.503 menos ventajoso (probable) que cuando la intensidad tecnológica es alta.

$e^{\hat{\beta}_8} = 1.524$, el Resultado de introducir un nuevo producto al mercado como éxito es 1.524 más ventajoso (probable) cuando el personal aumenta en uno.

Clasificar a una empresa

Si una empresa, de acuerdo a la base de datos, tiene como valores X1=6.25, X2=Prod. Nuevos, X3=Consumo industrial, X4=No, X5=Media, X6=20. Cómo se le clasificaría su Resultado de introducir un nuevo producto al mercado.

Ejemplo de Regresión Logística Multinomial (Archivo: PesoNiños)

Variable Dependiente

Peso_RN_1 : Peso del recién nacido-Clasificado

1. Bajo Peso

2. Peso Normal

3. Sobre Peso

Variables Independientes

Edad_Madre : Edad de la madre.

Num_Partos : Número de partos(Paridad).

Tipo_de_Parto : Tipo de parto (1=Cesárea, 2=Parto Normal)

Sexo_RN: Sexo del recién nacido (1=Femenino, 2=Masculino)

Talla_RN: Talla del recién nacido

Edad_Gestacional: Edad gestacional en semanas