CAPÍTULO 5
REGRESIÓN CON VARIABLES CUALITATIVAS

Edgar Acuña Fernández

Departamento de Matemáticas
Universidad de Puerto Rico
Recinto Universitario de Mayagüez

Regresión con variables predictoras cualitativas

- A veces algunas de las variables predictoras a considerar son cuantitativas y otras cualitativas.
- Ejemplos de variables cuantitativas son: años de experiencia, años de educación, edad, etc.
- Ejemplos de variables cualitativas son:

  sexo del empleado, estado civil, jerarquía del empleado, etc.

Tipos de Variables cualitativas

Variable “dummy” o variable binaria

- **Variables nominales**
- **Variables ordinales**

Nota:

En un problema de regresión debe haber por lo menos una variable predictora cuantitativa.

Si todas las variables predictoras fueran cualitativas entonces el problema se convierte en uno de **diseños experimentales.**

Regresión con una sola variable "Dummy"

- Consideremos un modelo de regresión con una sola variable cualitativa A y una variable cuantitativa X. Es decir,

$$Y=\beta_0+\beta_1 X+\beta_2 A + \varepsilon \ (*)$$

Consideremos los casos:

Si $A=0$, $Y= \beta_0 + \beta_1 X + \varepsilon$

Si $A=1$, $Y= (\beta_0 + \beta_2) + \beta_1 X + \varepsilon$

Es decir que el modelo (*) considera que las pendientes de ambas lineas son iguales.

El valor estimado de β_2 representa el cambio promedio en la variable de respuesta al cambiar el valor de la variable "dummy".

Si se desea comparar las pendientes de las línea de regresión de los dos grupos se puede usar una prueba de t similar a la prueba de comparación de dos medias y asumiendo que hay homogeneidad de varianza. También se puede usar una prueba de F parcial o probando la hipótesis Ho: $\beta 3=0$ en el siguiente modelo

$$Y=\beta 0+\beta 1A+\beta 2X+\beta 3AX +e$$

Cuando la hipótesis nula no es rechazada se concluye que la pendiente de regresion de ambos grupos son iguales. Si no hubiera igualdad de varianza de los dos grupos, habria que usar una prueba de t aproximada similar al problema de Behrens-Fisher. Aquí se usa una t con grados de libertad aproximados

Comparar Lineas de Regresión para Varios Grupos

Supongamos que se tiene una variable predictora continua X para explicar el comportamiento de Y en tres grupos. Surgen tres modelos de regresión:

i) $Y= \beta_{01}+\beta_{11}X+ \varepsilon$

ii) $Y=\beta_{02}+\beta_{12}X+ \varepsilon$

iii) $Y=\beta_{03}+\beta_{13}X+ \varepsilon$

Para relacionar las lineas de regresión hay que introducir 3 variables “dummy” para identificar los grupos G1, G2, y G3 y 3 variables adicionales Z1=G1X, Z2=G2X, y Z3=G3X.

Posibles modelos

a) Las lineas se intersecten en un punto cualquiera, ya que tendrían diferente intersecto y pendiente.

b) Las lineas sean paralelas (homogeneidad de pendientes).

c) Las lineas tengan el mismo intercepto pero distinta pendientes (homogeneidad de interceptos).

d) Las tres lineas coincidan.

Prueba de hipótesis

H_o: el modelo satisface b) o c) o d) versus

H_a: el modelo satisface a)

Se usa una prueba de F parcial dada por

$$F_m=[(SSE_m-SSE_a)/(gl_m-gl_a)]/[SSE_a/gl_a]$$

Donde *m*, representa los modelos b,c, o d.

grados de libertad.

La F parcial se distribuye como una F con (gl_m-gl_a, gl_a)

Regresión Logística

La variable de respuesta Y es una del tipo binario y que se tiene p variables predictoras x's (aleatorias)

Se tiene una muestra de tamaño $n=n_1+n_2$,

con n_1 observaciones de la clase C_1 y

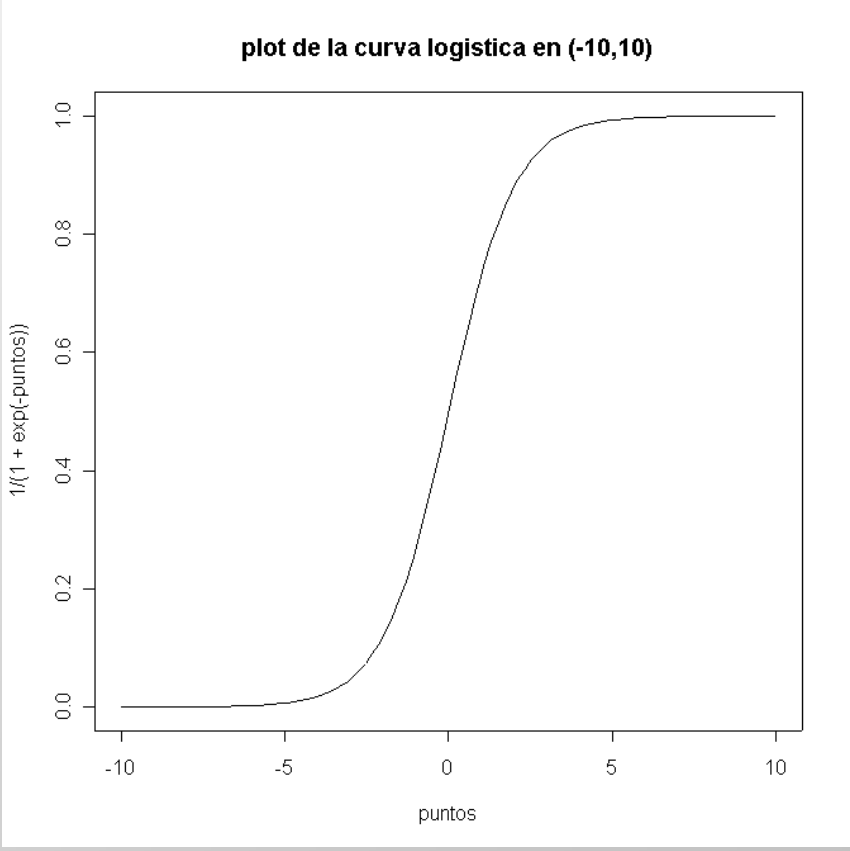
n_2 observaciones de la clase C_2 .

La variable predictora Y se define como 0 y 1 para cada clase.

La Curva Logística

Funcion de distribución logística

$$F(x) = \frac{1}{(1+e^{-x})}$$



Suposición en el Modelo Logístico

- Sea $f(\mathbf{x}/Ci)$ (i=1,2) la función de densidad del vector aleatorio p-dimensional $\mathbf{x}$ en la clase C_i, en el modelo logístico se asume que:

$$\log\left(\frac{f(\mathbf{x}/C_1)}{f\ (\mathbf{x}/C_2)}\right) = \alpha + \boldsymbol{\beta}'\mathbf{x}$$

Donde, β es un vector de p parámetros y α representa el intercepto.

Odds Ratio

Sea p=P(Y=1/**x**) la probabilidad a posteriori de que Y sea igual a 1 para un valor observado de **x**

Se define la razón de apuestas (***odds ratio***) como:

$$\frac{p}{1-p} = \frac{\frac{P\{Y=1\} f(\mathbf{x}/y=1)}{f(\mathbf{x})}}{\frac{P\{Y=0\} f(\mathbf{x}/y=0)}{f(\mathbf{x})}} = \frac{\pi_1 f\ (\mathbf{x}/C_1)}{\pi_2 f\ (\mathbf{x}/C_2)}$$

Donde: π_i representa la *probabilidad a priori* de que Y pertenezca a la clase C_i.

Transformación ***logit***

- Tomando logaritmos se tiene

$$\log(\frac{p}{1-p}) = \log(\frac{\pi_1}{\pi_2}) + \log\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}$$

- Luego con la suposicion se tiene que:

$$\log(\frac{p}{1-p}) = \alpha + \beta'\mathbf{x}$$

- Equivalentemente, $p = \dfrac{\exp(\alpha + \beta'\mathbf{x})}{1+\exp(\alpha + \beta'\mathbf{x})}$

Cumplimiento de la suposicion del modelo

Si las variables ***x*** en cada clase se distribuyen normalmente con medias $\boldsymbol{u_1}$, $\boldsymbol{u_2}$ y matríz de covarianza Σ entonces se satisface la suposición dado que:

$$\log\left(\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}\right) = (\boldsymbol{u_1}-\boldsymbol{u_2})' \, \boldsymbol{\Sigma^{1}}(\boldsymbol{x}-\boldsymbol{1/2}(\boldsymbol{u_1}+\boldsymbol{u_2}))$$

Donde

$$\alpha = -(\mathbf{u}_1 - \mathbf{u}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{u}_1 + \mathbf{u}_2)/2 \qquad \boldsymbol{\beta} = (\mathbf{u}_1 - \mathbf{u}_2)'\boldsymbol{\Sigma}^{-1}$$

La suposición también para otros tipos de distribuciones distintas de la normal multivariada tales como distribuciones de Bernoulli, y mezclas de éstas.

Estimación del modelo logístico

Método de Máxima Verosimilitud

Dada una observación **x**, las probabilidades de que ésta pertenezca a las clases C1 y C2 son :

$$P(C_1 / \mathbf{x}) = \frac{\exp(\alpha + \beta' \mathbf{x})}{1 + \exp(\alpha + \beta' \mathbf{x})} \qquad P(C_2 / \mathbf{x}) = 1 - P(C_1 / \mathbf{x}) = \frac{1}{1 + \exp(\alpha + \beta' \mathbf{x})}$$

Considerando una muestra de tamaño $n=n_1+n_2$ y un parámetro binomial *p* igual a $\exp(\alpha + \boldsymbol{\beta}' \mathbf{x}) / (1 + \exp(\alpha + \boldsymbol{\beta}' \mathbf{x}))$ la función de verosimilitud es de la forma

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n_1} \frac{\exp(a + \mathbf{x_i}' \boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{x_i}' \boldsymbol{\beta})} \cdot \prod_{j=n_1+1}^{n} \frac{1}{1 + \exp(\alpha + \mathbf{x_j}' \boldsymbol{\beta})}$$

La solucion de la equacion de verosimilitud es solo numerica

Otra forma de estimar los parámetros α y β

Consideramos la regresión lineal múltiple de $\text{logit}(\hat{p})$ versus $x_1, x_2, \ldots, x_p$

luego su varianza será aproximadamente:

$$Var[\ln(\frac{\hat{p}}{1-\hat{p}}] \cong [\frac{1}{p(1-p)}]^2 \frac{p(1-p)}{n_1} = \frac{1}{n_1 p(1-p)}$$

Asi se llega a un problema donde la varianza no es constante y se puede usar mínimos cuadrados ponderados con pesos $w_i(\mathbf{x}) = n_1 \hat{p}(\mathbf{x})(1- \hat{p}(\mathbf{x}))$ para estimar los parámetros α y β del modelo logístico.

Medidas de Confiabilidad del Modelo

Cuantifican el nivel de ajuste del modelo al conjunto de datos:

- La Devianza
- El Pseudo-R^2
- El Criterio de Información de Akaike (AIC)
- La Prueba de Bondad de Ajuste de Hosmer-Lemeshov.

La Devianza

Es similar a la suma de cuadrados del error de la regresión lineal.
Si la variable de respuesta Y no está grupada se tiene que:

$$D = -2\{\sum_{i:y_i=1}^{n} \log(\hat{p}_i) + \sum_{i:y_i=0}^{n} \log(1-\hat{p}_i)\}$$

D tiene una distribución Ji-Cuadrado con (n-p-1)gl.
Si D es mayor que una Ji-Cuadrado con (n-p-1)gl
para un nivel de significación dado entonces el modelo logístico no es confiable.

El Pseudo-R^2

- Es similar al R^2 de la regresión lineal se define por:

$$Pseudo - R^2 = (1 - \frac{Devianza}{Devianza.Nula})100\%$$

Donde la *DevianzaNula* es la devianza considerando solamente el intercepto y se distribuye como una *Ji-Cuadrado* con (*n*-1)*gl*. Para hallar la *DevianzaNula* se hace una regresión logística considerando que hay una sola variable predictora cuyos valores son todos unos.

El Criterio de Información de Akaike (AIC)

Determina si un modelo es mejor que otro.

Esta dado por:

$$AIC=D+2(p+1)$$

Donde, *p* es el número de variables predictoras.

Un modelo es mejor que otro si su AIC es más pequeño.

La Prueba de Bondad de Ajuste de Hosmer-Lemeshov.

Se aplica cuando los datos son dados en forma agrupada y se define por

$$C = \sum_{i=1}^{g} \frac{(O_i - n'_i \, \bar{p}_i)^2}{n'_i \, \bar{p}_i (1 - \bar{p}_i)}$$

Donde g es el número de grupos, n'_i es el número de observaciones en el *i-ésimo* grupo

O_i es la suma de las y's en el *i-ésimo* grupo y $\bar{p}_i$ es el promedio de las p_i en el *i-ésimo* grupo.

Estadisticas Influenciales para regresión logística

- ***Residuales de Pearson*** es similar al residual estudentizado usado en regresión lineal

$$r_i = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)}}$$

y_i representa el número de veces que y=1 entre las m_i repeticiones de X_i , de lo contrario m_i=1 para todo i.

- **Residuales de Devianza** Esta dado por:

si y_i=0 y si y_i=1

Una observación será anormal $D_i = -\sqrt{2|\log(1-\hat{p}_i)|}$ $D_i = \sqrt{2|\log(\hat{p}_i)|}$ de devianza es mayor que 2 en valor absoluto, entonces la observación correspondiente es anormal.

Uso de la regresión logística en Clasificación:

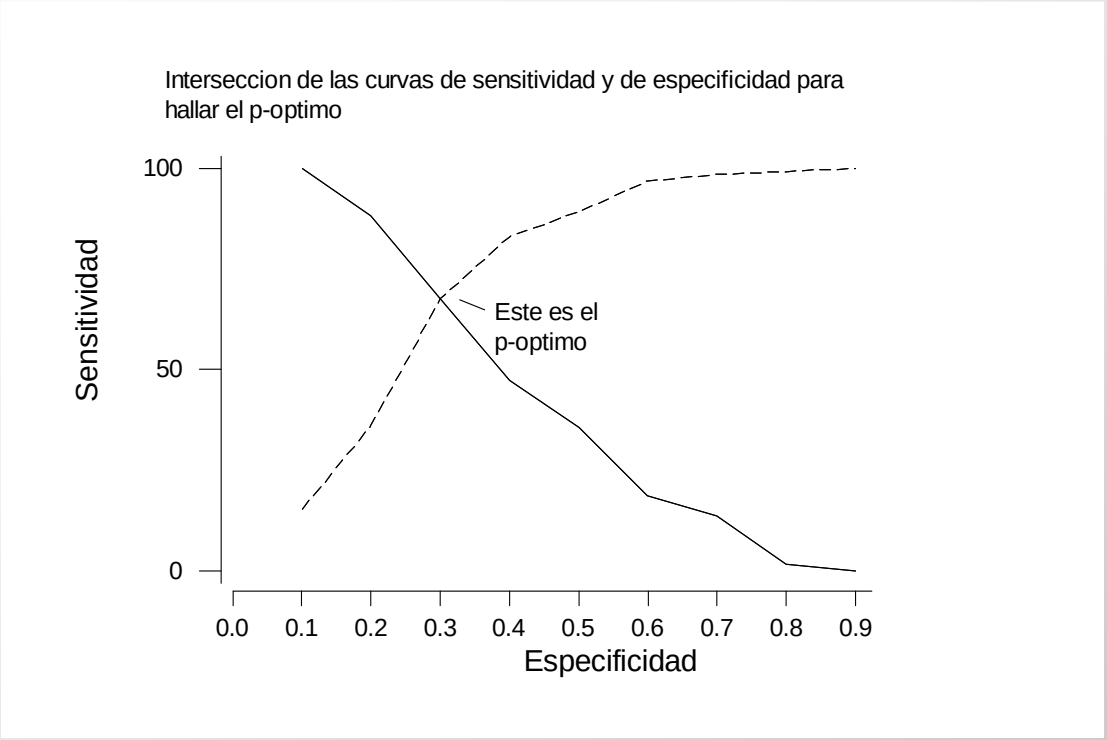
La forma mas facil de discriminar es considerar que si p>0.5 la obervación pertenece a la clase que uno está interesado.

Otros metodos son:

i) *Plotear el porcentaje de obervaciones que poseen el evento que han sido correctamente clasificadas (Sensitividad) versus distintos niveles de probabilidad* y el porcentajes de observaciones de la otra clase que han sido correctamente clasificadas (especifidad) versus los mismos niveles de probabilidad anteriormente usados, en la misma gráfica.

La probabilidad que se usará para clasificar las observaciones se obtienen **intersectando las dos curvas.**

Ejemplo



Las gráficas de los dos métodos aparecen en la siguiente figuras y embos caso el p-óptimo a usarse es p=0.3

La curva ROC

ii) Usar la curva ROC (Receiver Operating Characteristic Curva).

En este caso se grafica

la sensitividad versus (1-especifidad)100%,

y se coge como el *p* ideal aquel que **está más cerca a la esquina superior izquierda**, osea al punto (100,0).

Ejemplo de curva ROC