MAGISTER EN ESTADÍSTICA

REGRESIÓN LOGÍSTICA

Agenda

Introducción

La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.

Introducción

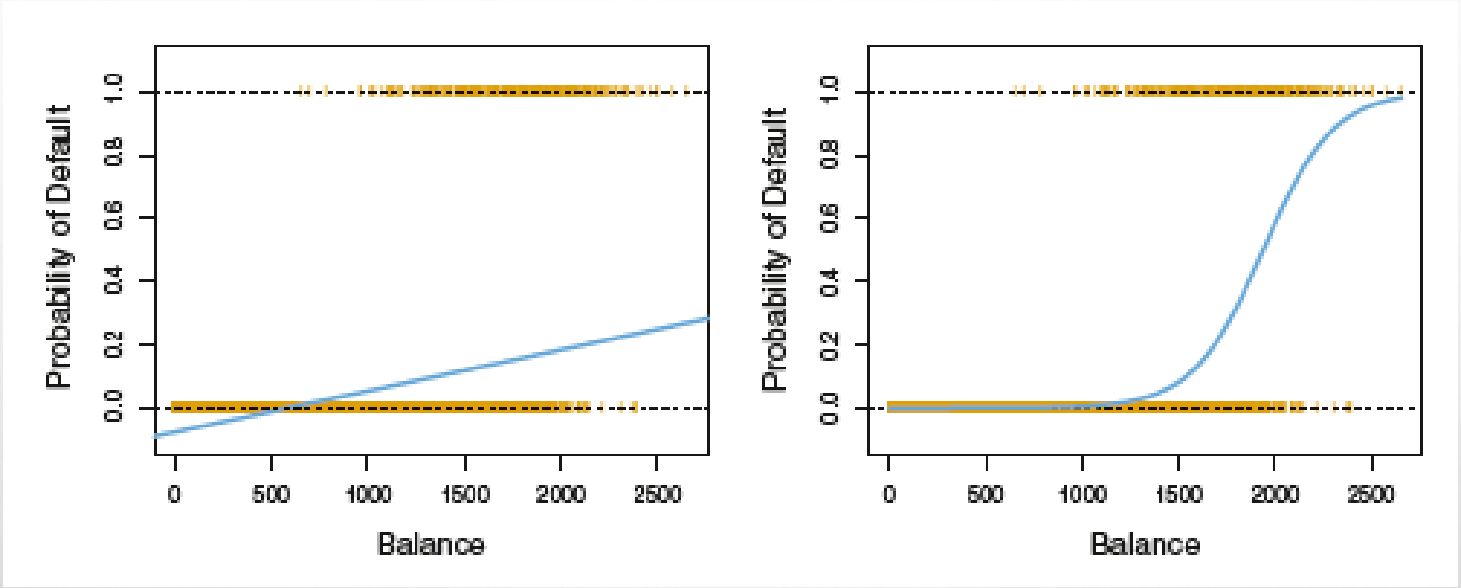

Figure: Probabilidad estimada de impago utilizando regresión lineal (izquierda) y Probabilidad de impago mediante regresión logística (Derecha).

Regresión Logística

El modelo de regresión logística permite estimar la probabilidad de que un individuo pertenezca o no pertenezca a una población o categoría (siendo ésta siempre binaria), que depende de los valores de ciertas covariables.

Regresión Logística

Supongamos que un suceso (o evento) de interés A puede presentarse o no en cada uno de los individuos de cierta población. Consideremos la variable binaria y que toma los valores:

$$y = 1 \text{ si A se presenta}, \quad y = 0 \text{ si A no se presenta}$$

Regresión Logística

Verosímilitud

Si la probabilidad de A no depende de otras variables, indicando $P(A) = p$ la verosimilitud de una única observación y es

$$L(p) = p^y(1-p)^{1-y}.$$

Donde $L(p) = p$ si $y = 1$, $\quad L(p) = 1 - p$ si $y = 0$.

Regresión Logística

Propiedades

Si se realizan n pruebas independientes y observamos $y_1, \ldots, y_n$, la verosimilitud es

$$L(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^k(1-p)^{n-k}.$$

Siendo $k = \sum_{i=1}^{n} y_i$, la frecuencia absoluta de A en las n pruebas. Para estimar p resolvemos la ecuación de verosimilitud

$$\frac{\partial}{\partial p} \ln\, L(p) = 0$$

Se tiene que la solución es $\widehat{p} = k/n$, la frecuencia relativa del suceso A. La distribución asintótica de $\widehat{p}$ es normal $N(p, p(1-p)/n)$.

Regresión Logística

Modelo

Supongamos que la probabilidad p depende de los valores de ciertas variables $X_1, \ldots, X_p$. Es decir, si $x = (x_1, \ldots, x_p)'$ son las observaciones de un cierto individuo, entonces la probabilidad de que suceda A dado $\mathbf{x}$ es $p(y = 1|\mathbf{x})$. Se indicará esta probabilidad por $p(\mathbf{x})$. La probabilidad contraria de que A no suceda dado $\mathbf{x}$ será $p(y = 0|\mathbf{x}) = 1 - p(\mathbf{x})$.

$$\ln\left[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta_0 + \boldsymbol{\beta}'\mathbf{x},$$

siendo $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ parámetros de regresión. El modelo es equivalente a suponer las siguientes probabilidades para A y su complemento, ambas en función de $\mathbf{x}$

$$p(\mathbf{x}) = \frac{e^{\beta_0+\beta'\mathbf{x}}}{1 + e^{\beta_0+\beta'\mathbf{x}}}, \quad 1 - p(\mathbf{x}) = \frac{1}{1 + e^{\beta_0+\beta'\mathbf{x}}}$$

Regresión Logística

Selección de variable

Comúnmente se tienen base de datos con una gran cantidad de variables, de las cuales no todas son significativas para la predicción, es por esto que se tienen distintos métodos para hacer selección de variables para obtener la mejor predicción.

- **Backward**: se comienza por considerar el modelo con todas las variables disponibles y se van eliminando del modelo de una en una, según su capacidad explicativa. Es decir, la primera variable eliminada es aquella que presenta un menor coeficiente de correlación parcial con la variable dependiente.
- **Forward**: se comienza por considerar que el modelo no contiene ninguna variable explicativa y se añade como la primera de ellas a la que presente un mayor coeficiente de correlación parcial con la variable dependiente.
- **Stepwise**: es uno de los métodos más aplicados y consiste en una combinación de los dos anteriores. En el primer paso se procede como en el método **Forward** pero a diferencia de éste en el que cuando la variable entra al modelo ya no vuelve a salir en el procedimiento.

Regresión Logística

Supuestos

- Independencia: las observaciones tienen que ser independientes unas de otras.
- Relación lineal entre el logaritmo natural de odds y la variable continua: patrones en forma de U son una clara violación de esta condición.
- La regresión logística no precisa de una distribución normal de la variable continua independiente.
- Número de observaciones: no existe una norma establecida al respecto, pero se recomienda entre 50 a 100 observaciones.

Regresión Logística

Supuestos

Supongamos tenemos $x_1, \ldots, x_n$ variables que nos interesaría verificar si son significativas o no, para ello debemos contrastar si los coeficientes de las variables son 0 o no:

$$H_0 : \beta = 0 \quad vs \quad \beta \neq 0$$

El contraste anterior se construye considerando que el estimador es β.

$$\hat{\beta} = \hat{\beta}_0, \ldots, \hat{\beta}_n$$

Si se tiene que n es grande, la distribución de $\hat{\beta}$ tiene una distribución normal.

$$\hat{\beta} \sim N(\beta, \widehat{VAR}(\hat{\beta}))$$

Regresión Logística

Supuestos

A partir de la distribución normal se tiene que el estadístico de contraste para la prueba de hipótesis

$$H_0 : \beta = 0 \quad vs \quad \beta \neq 0$$

viene dada por:

$$Z = \frac{\hat{\beta}}{ee(\hat{\beta})}$$

Bajo la hipótesis H_0, $Z \sim N(0,1)$. Se rechaza H_0 si $|Z| > Z_{\frac{\alpha}{2}}$ donde α es el nivel de significación del contraste y $Z_{\frac{\alpha}{2}}$ el cuantil $1 - \alpha/2$ de la distribución $N(0,1)$.

Criterio de clasificación

Una de las principales aplicaciones de un modelo de regresión logística es clasificar la variable cualitativa en función de valor que tome el predictor. Para conseguir esta clasificación, es necesario establecer un threshold de probabilidad a partir de la cual se considera que la variable pertenece a uno de los niveles. Por ejemplo, se puede asignar una observación al grupo 1 si $\hat{p}(Y = 1|X) > 0.5$ y al grupo 0 si de lo contrario.