

Logit y otros análisis

Fernando A. Crespo R.

July 12, 2023



Índice

5. Complementos a la Regresión Múltiple Lineal

5.1 Modelo Logit

5.2 Análisis de la Varianza

5.3 Análisis de la Varianza Multivariante MANOVA

5.1 Modelo Logit

- ▶ Modelo Logit o logístico, se aplica cuando uno quiere predecir, o usar como variables predictivas, tasas de crecimiento: introducción de nuevos productos, ciclo de vida del producto.
- ▶ $y_t \in \{0, 1\}$.
- ▶ Se propone la forma logística:

$$y_t = \frac{1}{1 + e^{-(\mathbf{x}'_t\beta + \delta t + \epsilon_t)}}, \quad (1)$$

- ▶ Se calcula como:

$$\text{logit}(y_t) = \ln \left(\frac{y_t}{1 - y_t} \right) = \mathbf{x}'_t\beta + \delta t + \epsilon_t, \quad (2)$$

5.1 Modelo Logit

- ▶ Veamos el efecto de función logit, sobre las probabilidades.
- ▶ Veamos un ejemplo, Ver el efecto de la anestesia en función de la concentración, del punto de vista que la persona se mueve o no se mueve.
- ▶ Para el ejemplo, la logit empírica, si x son los pacientes que no se mueven es $\log\left(\frac{x+0.5}{n-x+0.5}\right)$.

5.1 Modelo Logit

- ▶ Veamos un caso multivariado. La idea es ver una distribución de ranas de Southern Corroboree, en las Snowy Mountains de New South Wales. La variable es presencia o ausencia de la rana.
- ▶ Variables: easting (punto de referencia este), northing (punto de referencia norte), altitude (altitud en metros), distance (distancia a la población más cercana), NoOfPools (número de potenciales piscinas de cría), NoOfSites (número de potenciales sitios de cría a 2km de radio), avrain (lluvia promedio del período de primavera), meanmin (media de la temperatura mínima en primavera), meanmax (media de la temperatura máxima en primavera).
- ▶ Primer paso, ver la relación entre las variables explicativas.

5.1 Modelo Logit

- ▶ Como se quiere ver las variables que influyen, se sacarán las variables de posición, para evitar explicar por localización.
- ▶ La relación de las dos primeras no es clara, se le aplicará logaritmo a la distancia y raíz cuadrada a NoOfPools. En NoOfSites hay valores cero, se aplica $\log(\text{NoOfSites} + 1)$.
- ▶ arreglar los datos muy correlacionados. Mejor usar $\text{meanmax} + \text{meanmin}$ ó $\text{meanmax} - \text{meanmin}$, que las variables separadas.
- ▶ La función logit:

$$f(x) = \log \left(\frac{x}{1-x} \right), 0 < x < 1. \quad (3)$$

- ▶ La función inversa:

$$g(u) = \frac{e^u}{1 + e^u}. \quad (4)$$

5.1 Modelo Logit

- ▶ Al modelo hay que ajustar la respuesta. Usando `fitted()` o `predict()`.
- ▶ Se puede ver el aporte de las variables explicativas al modelo.
- ▶ Hay que ver la validación cruzada. Donde el programa muestrea los datos. Usa una parte de los datos para el modelo y otra para validar. Se trata de usar todos los datos, al menos para construir el modelo, y para validarlo.

5.1 Modelo Logit

- Para validar el modelo se usa la matriz de confusión:

Valor Predicción

		Valor Predicción		
		p	n	total
actual value	p'	Verdadero Positivo	Falso Negativo	p'
	n'	Falso Positivo	Verdadero Negativo	N'
total		P	N	

5.1 Modelo Logit

- ▶ En Ciencia de Datos, el conjunto de datos es particionado en dos grupos, generalmente en una regla del 70 u 80 % para entrenar el modelo, y los restantes datos para testear el modelo.
- ▶ De esa manera se construyen dos tablas de confusión, una para datos de entrenamiento y otra para los datos de testeo.

5.1 Modelo Logit

- El cálculo que se obtiene de aplicar la exponencial a los ponderadores del modelo logit, se llaman OR indican el nivel de riesgo para ser predicho como valor 1.

5.1 Modelo Logit

- ▶ Se puede hacer un modelo Logit cuando hay variables categóricas.
- ▶ Hay que desarrollar un tabla con múltiples entradas para los datos, veremos el ejemplo `capitulo5_parte2.R`
- ▶ Usaremos `UCBAdmissions`, que es una tabla 2X2X6 con variables: Admit (Admitted-Rejected) X Gender (Male-Female) X Dept (A,B,C,D,E,F).

5.2 Análisis de la Varianza

- ▶ Los datos tienen variables que identifican sus características de manera única, como tratamientos, pertenencia, etc., a ello le llaman factores. Deben ser excluyentes entre sí.
- ▶ Ejemplo: sugar
- ▶ Primero vemos si el modelo es significativo, es decir si hay diferencias globales.
- ▶ con el Test de Tukey, vemos si hay diferencia entre factores.

5.2 Análisis de la Varianza

- ▶ Un ejemplo interesante es ver la evaluación de gusto.
- ▶ En este caso se evaluó cuatro tipos de manzanas, y a cada panelista se le omitió algún tipo de manzana, para que diferenciara el gusto de las otras.
- ▶ fueron 20 panelistas, y evaluaban entre 0 (no les gustó) a 150 (les gustó mucho).

5.3 Análisis de la Varianza Multivariante MANOVA

- ▶ La idea es hacer análisis de la varianza con múltiples características y en múltiples etapas.
- ▶ Por ejemplo perfiles de enfermedad con sus variables medidas y sus variables diferenciadoras en distintos momentos del tiempo.
- ▶ Supone desviación estándar homogénea y distribución normal en los datos a medir.