Del Dato
al Conocimiento

Modelos Lineales Generalizados

Antonio Pita Lozano

Máster en Data Science

Índice

Del Dato
al Conocimiento

Modelización Lineales Generalizados

Los modelos lineales generalizados son técnicas de modelización estadística que se utilizan cuando la variable dependiente se identifica con una distribución conocida incluida en la familia exponencial. Son una generalización de la regresión lineal.

Principales técnicas:

- Regresión Bernoulli Logística
- Regresión Bernoulli Probit
- Regresión Poisson
- Regresión Beta
- Regresión Gamma
- Regresión Geométrica
- Regresión Chi-cuadrado
- Regresión Exponencial

Regresión Bernoulli Logística

Del Dato al Conocimiento

Objetivo

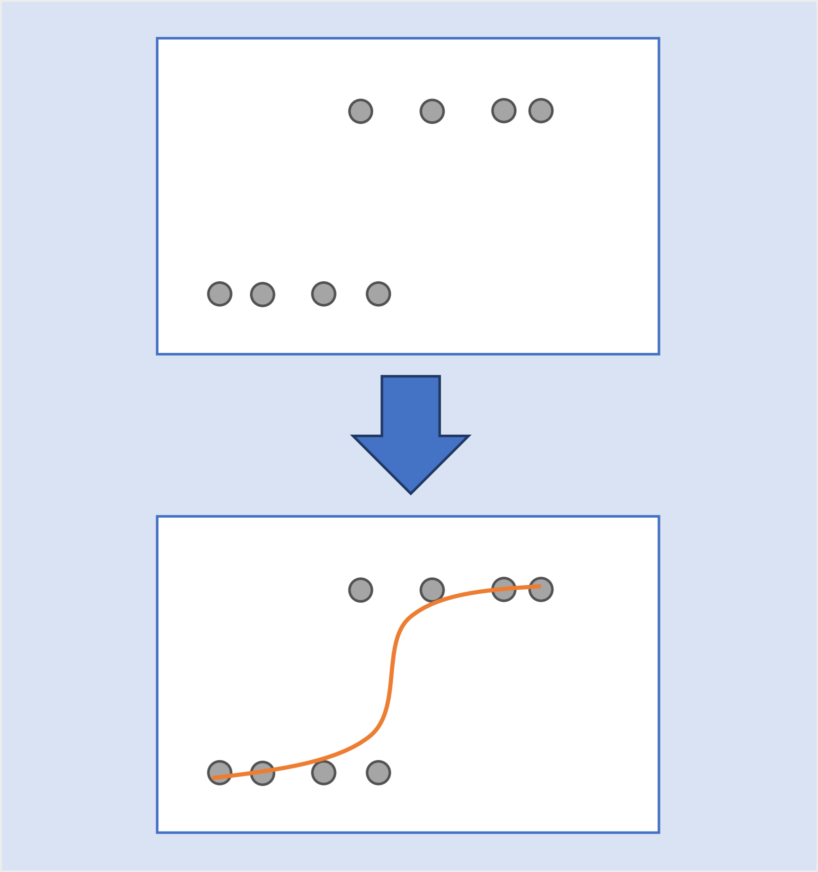
Estimar la relación entre una variable dependiente (variable explicada que sigue una distribución binaria Bernoulli) y varias variables independientes (variables explicativas) mediante una expresión lineal en coeficientes y la función link logit.

Desarrollo

Para la estimación de una relación líneas es necesario establecer el modelo a estimar, que será una combinación lineal de los regresores y una función link logit, también llamada función sigmoide:

$$P(y=1|x)=\frac{1}{1+e^{-z}} \text{ donde } z=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\varepsilon$$

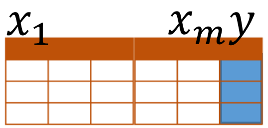
La estimación se realiza utilizando el estimador máximo verosímil y las estimaciones se obtienen mediante un algoritmo recursivo de mínimos cuadrados ponderados o la técnica de Newton-Raphson



1

Regresión Bernoulli Logística: Componentes

Datos



Modelo

$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$ donde P(Y=1|x)=$\frac{1}{1+e^{-z}}$

Función de coste

$g(\beta_0, \beta_1, \ldots, \beta_m) = \ln(L(\beta_0, \beta_1, \ldots, \beta_m))$

Algoritmo estimador

Estimador Máximo Verosímil mediante al algoritmo recursivo de Newton-Raphson

Regresión Bernoulli Logística: Interpretación Coeficientes

Medidas de efecto de Riesgos

- Odds es una medida de

$$O(X) = \frac{P(Y = 1|x)}{1 - P(Y = 1|x)}$$

- Odds Ratio o razón de monios es una medida del efecto entre dos situaciones

$$OR(X, X') = \frac{O(X)}{O(X')}$$

- Risk Ratio es una medida de la ocurrencia ante un evento en comparación con la ausencia del evento.

$$RR(X, X') = \frac{R(X)}{R(X')}$$

	Afectado	No afectado	
Expuesto	EA	ENA	TE
No Expuesto	NEA	NENA	TNE
	TA	TNA	

	Afectado	No afectado	
Expuesto	70	30	100
No Expuesto	50	150	200
	120	180	

$$O(Expuesto) = \frac{70/100}{30/100} = 2{,}33\ veces$$

$$O(NoExpuesto) = \frac{50/200}{150/200} = 0{,}33\ veces$$

$$OR(E, NE) = \frac{2{,}33}{0{,}33} = 7\ veces$$

$$RR(E, NE) = \frac{70/100}{50/200} = 2{,}8\ veces$$

Coeficiente de la regresión logística

- Coeficiente o log-odds ratio

$$\beta_i = \log(OR(X_i, X)) = \log(\frac{O(X_i)}{O(X)})$$

- Exponencial del coeficiente u odd-ratio

$$e^{\beta_i} = OR(X_i, X) = \frac{O(X_i)}{O(X)}$$

Regresión Bernoulli Logística

Regresión Bernoulli Probit

KSCHOOL

Del Dato al Conocimiento

Objetivo

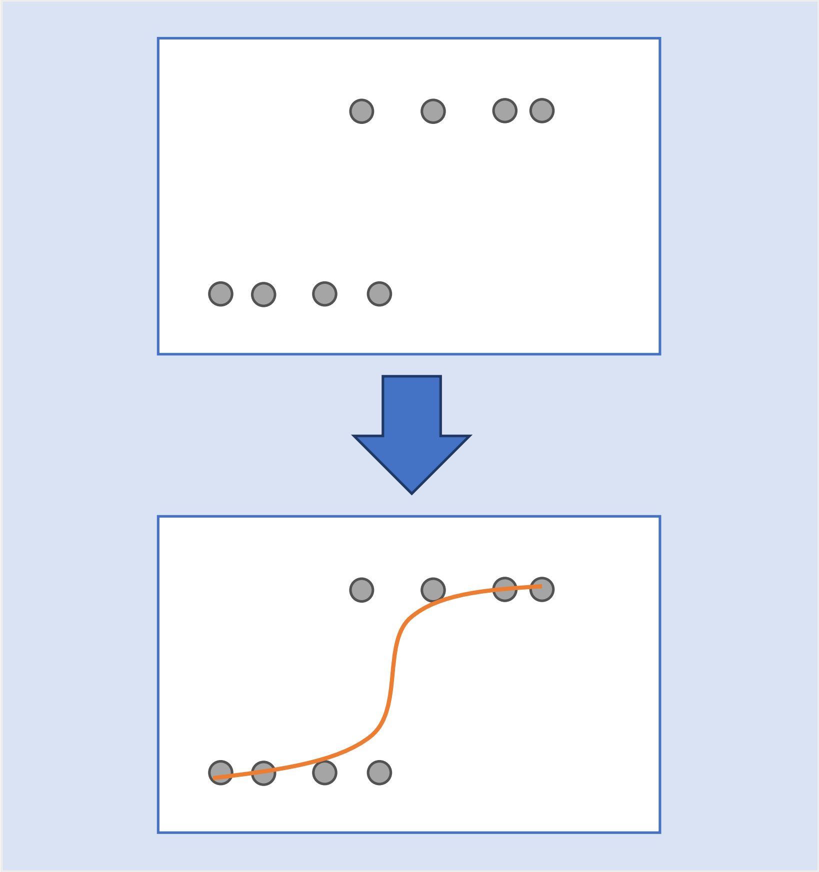
Estimar la relación entre una variable dependiente (variable explicada que sigue una distribución binaria Bernoulli) y varias variables independientes (variables explicativas) mediante una expresión lineal en coeficientes y la función link gausiana.

Desarrollo

Para la estimación de una relación líneas es necesario establecer el modelo a estimar, que será una combinación lineal de los regresores y una función link probit, que es precisamente la función de distribución de la distribución normal estandar:

$$P(y=1|x)= \theta(z) \text{ donde } z= \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

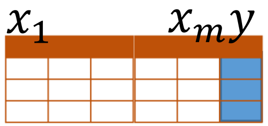
La estimación se realiza utilizando el estimador máximo verosímil y las estimaciones se obtienen utilizando algoritmos de optimización numéricos que convergen al óptimo global dado que la función es cóncava (problema de optimización conveza) y siempre tiene solución óptima.



Regresión Bernoulli Probit: Componentes

Del Dato al Conocimiento

Datos



Modelo

$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$ donde P(Y=1|x)=$\theta(z)$

Función de coste

$\mathrm{g}(\beta_0, \beta_1, \dots, \beta_m) = \ln(L(\beta_0, \beta_1, \dots, \beta_m))$

Algoritmo estimador

Diversos algoritmos de optimización convexa como descenso del gradiente

No hay interpretación de coeficientes sencilla

Regresión Bernoulli Probit

Regresión Poisson

Objetivo

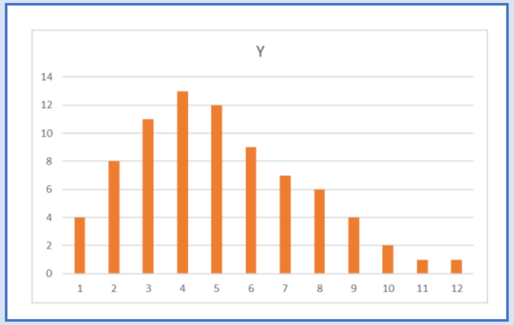
Estimar la relación entre una variable dependiente (variable explicada que sigue una distribución discreta Poisson) y varias variables independientes (variables explicativas) mediante una expresión lineal en coeficientes y la función link logaritmo.

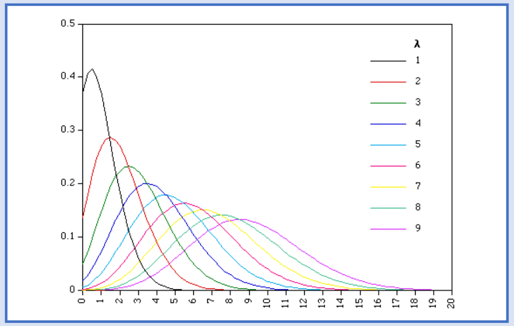
Desarrollo

Para la estimación de una relación líneas es necesario establecer el modelo a estimar, que será una combinación lineal de los regresores y una función link logit, también llamada función sigmoide:

$$E(y|x)= e^{z} \text{ donde } \qquad z= \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

La estimación se realiza utilizando el estimador máximo verosímil y las estimaciones se obtienen utilizando algoritmos de optimización numéricos que convergen al óptimo global dado que la función es cóncava (problema de optimización conveza) y siempre tiene solución óptima.



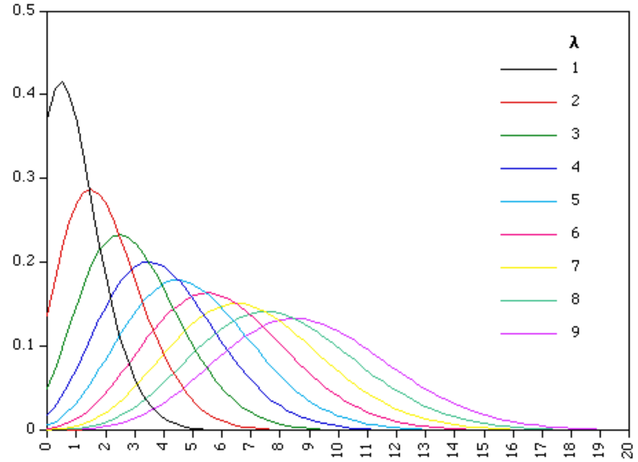


Regresión Poisson: Componentes

	Datos	$x_1 \quad x_m \, y$
	Modelo	$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$ donde E(Y|x)=e^z
	Función de coste	$\mathrm{g}(\beta_0, \beta_1, \ldots, \beta_m) = \ln(L(\beta_0, \beta_1, \ldots, \beta_m))$
	Algoritmo estimador	Diversos algoritmos de optimización convexa como descenso del gradiente

2 Regresión Poisson: Interpretación

Función de densidad



Output de la regresión Poisson

$$E(Y|x) = e^{\beta_0+\beta_1 x_1+\cdots+\beta_m x_m} = \lambda$$

Coeficientes de la regresión Poisson

- X Variable binaria → $\exp(\beta_i)$ es la razón de aumento de la media condicional

$$\exp(\beta_i) = \frac{P(Y|X_i = 1)}{P(Y|X_i = 0)}$$

- X Variable continua → β_i es la semielasticidad

Regresión Poisson

Del Dato al Conocimiento

el rincón de Antonio Pita

https://antoniopita.blog

Del Dato al Conocimiento

Introducción a la Modelización Estadística

Antonio Pita Lozano

Máster en Data Science

https://www.linkedin.com/in/antoniopitalozano/

@anto_pita