Auxiliar 9

MNL y MLE

Daniel Olcay

IN4402

21 de octubre de 2014

Índice

- Modelos no lineales
- Probabilidad lineal
- Probit
- Logit
- Máxima verosimilitud

Introducción

- Hasta ahora hemos visto problemas en los que nos interesa modelar una variable continua, sin embargo existe una infinidad de aplicaciones en las que nos gustaría modelar una variable aleatoria discreta
  1. Cliente permanece o se va de una compañía.
  2. País cae en *default* o no.
  3. Individuo es pobre o no.
  4. etc
- ∞'s aplicaciones!!!

Modelo General

- Bajo las suposiciones clasicas, el modelo lineal es $Y = X\beta + u$, y nos interesa $\mathbb{E}(Y|X) = X\beta$
- Supongamos sin perdida de generalidad que $X\beta = \beta_0 + \beta_1 X_i$
- El modeo general esta dado por

$$\mathbb{E}(Y|X) = F(X\beta)$$

Donde F es una función a especificar

Modelo de probabilidad lineal

- El modelo de probabilidad lineal es

$$\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

- Si Y es binaria, se tiene que

$$\mathbb{E}(Y_i|X_i) = 1 \cdot \mathbb{P}(Y_i = 1|X_i) + 0 \cdot \mathbb{P}(Y_i = 0|X_i)$$

- Por lo que el problema queda especificado por

$$\mathbb{P}(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$$

Ventajas/Desventajas Modelo de probabilidad lineal

- El modelo de probabilidad lineal es fácil de estimar.
- De igual modo, la inferencia resulta sencilla
- Sin embargo, posee un problema a nivel de predicción. Nadie asegura que el valor predicho pertenezca al intervalo $[0, 1]$!!!

Probit

- El modelo Probit surge de asumir que F es una distribución normal
- Así, el modelo queda especificado de la siguiente forma

$$\mathbb{P}(Y_i = 1|X_i) = \Phi(\beta_0 + \beta_1 X_i)$$

- Ventajas:
  - Función de distribución asegura consistencia en cuanto a dominio.
  - Distribución normal tiene una reputación amplia en estadísticas.
  - Probabilidades tabuladas
- Desventajas:
  - Problema de maximización asociado para estimación no tiene solución explicita
  - Recurrir a aproximaciones numéricas.

Logit

- El modelo logit surge de asumir F siguiendo una distribución logística.
- Distribución logística

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

- Así, el modelo queda especificado como

$$\mathbb{P}(Y_i = 1 | X_i) = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_i}}$$

- Posee ventajas computacionales respecto al Probit, sin embargo en la práctica poseen resultados similares.

Estimación

- Estimación mediante mínimos cuadrados posee desventajas respecto a otros métodos de estimación en el contexto de MNL.
- A modo de ejemplo, modelo Probit se estima mediante

$$\max_{b_0,b_1} \sum_{i=1}^{n} [Y_i - \Phi(b_o + b_1 X_i)]^2$$

- No hay solución cerrada dicho problema
- Usar métodos numéricos: Gradiente, Newton-Rhapson, Quasi-Newton, etc.
- Hay métodos más eficientes...

Máxima verosimilitud

- Idea intuitiva: Maximizar la probabilidad de observar lo que se esta observando.
- Sea $\{y_1, y_2, \ldots, y_n\}$ una muestra de valores. Sea además, θ un vector de parámetros de interés.
- Sea $f(y|\theta)$ la densidad conjunta asociada. Esto es, la función que determina la probabilidad de ocurrencia de y dado el parámetro θ.
- La **verosimilitud** estará dada por

$$L(\theta|y) = f(y|\theta)$$

Máxima verosimilitud (2)

- Al maximizar $L(\theta|y)$ respecto a θ se obtienen los parámetros que **hacen mas probable la ocurrencia de los datos**
- El estimador máximo verosimil $\hat{\theta}_{MLE}$ será aquel que maximize la probabilidad de ocurrencia de la muestra observada.
- Usualmente, maximizamos el logaritmo de la función verosimilitud, la cual apodamos como **log-verosimilitud**

$$\hat{\theta}_{MLE} = arg \max_{\theta} \ln(L(\theta|y))$$

- Recordemos que $L(\theta|y) = f(y|\theta)$...¿Cómo calculamos la densidad conjunta?

Distribución conjunta - supuestos

- Estamos interesados en la probabilidad conjunta de **todos los datos**. Considerando una muestra de n observaciones

$$L(\theta|y) = f(y_1, y_2, \ldots, y_n|\theta)$$

- Asumiremos que la muestra $\{y_i\}_{i=1}^{n}$ es i.i.d.
- Independencia: Asegura que la densidad conjunta es la multiplicación de las marginales
- Identicamente distribuidos: Los parámetros θ son los mismos para toda la población.

$$L(\theta|y) = f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\theta)$$

Log-verosimilitud y estimación

- Finalmente, aplicando logaritmo la log-verosimilitud queda

$$\ln(L(\theta|y)) = \ln\left(\prod_{i=1}^{n} f(y_i|\theta)\right) = \sum_{i=1}^{n} \ln(f(y_i|\theta))$$

- Estimación
  1. Plantear verosimilitud
  2. Calcular log-verosimilitud
  3. Derivar la log-verosimilitud respecto a los parámetros de interes
  4. Despeje

Distribución asintótica MLE

- Dado que el estimador máximo verosimil es una variable aleatoria, es posible realizar inferencia sobre el.
- Bajo ciertas condiciones teóricas y usando el TCL, CMT y Slutsky, es posible mostrar que la distribución asintótica del estimador MLE es

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \sim N\left(0, \mathbb{I}(\theta)^{-1}\right)$$

- $\mathbb{I}(\theta)$ se conoce como **matriz de información** y se calcula como

$$\mathbb{I}(\theta) = -\mathbb{E}(H(\ln(L(\theta|y))))$$

  En palabras: la esperanza del hessiano de la log-verosimilitud!
- Inferencia: Intervalos de confianza usuales