

Estimação por Máxima Verossimilhança

Helson Gomes de Souza

Dezembro de 2020

A função de verossimilhança

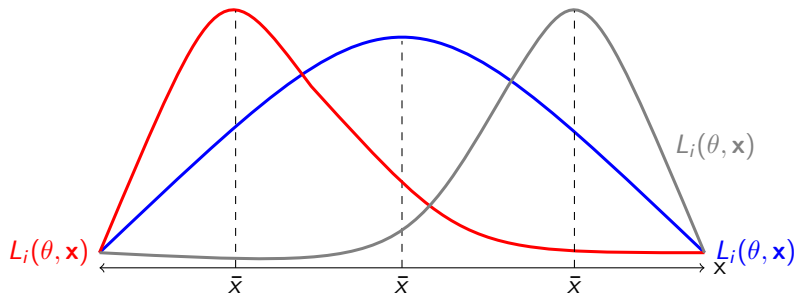
- Considere $(x_1, x_2, \dots, x_N)$ como sendo uma amostra aleatória que tem uma distribuição com densidade dada por $f(x, \theta)$, em que $\theta \in \Theta$ é um vetor $p - 1$ de parâmetros desconhecidos e $\Theta \in \mathbb{R}^p$ é o espaço vetorial que agrupa os parâmetros.
- A função de verossimilhança corresponde à densidade de probabilidade conjunta de $(x_1, x_2, \dots, x_N)$, vista, entretanto como uma função dos parâmetros desconhecidos θ .
- Formalmente, esta função pode ser descrita como $L_i(x_1, x_2, \dots, x_N)$, ou simplesmente:

$$L_i(\theta, \mathbf{x}) = f(x_1, x_2, \dots, x_N, \theta) \quad (1)$$

- Em que $\mathbf{x} = (x_1, x_2, \dots, x_N)'$.

A função de verossimilhança

- Note que o ponto em que a função de verossimilhança está sendo maximizada corresponde à média da distribuição.
- Como uma estimação paramétrica busca obter o vetor θ de parâmetros desconhecidos que represente uma relação por meio da média, é conveniente considerar que os valores do vetor θ seriam aqueles que atribuem um valor máximo para a função de verossimilhança.



A função de verossimilhança

- Contudo, se a distribuição é muito assimétrica, a média não é uma boa medida de inferência. Com isso, os estimadores de máxima verossimilhança podem apresentar resultados pouco representativos em distribuições assimétricas.
- Para encontrar a estimativa de máxima verossimilhança é preciso encontrar o valor máximo da função de verossimilhança. Visto que $\ln(x)$ é uma função crescente de x , a aplicação do logaritmo facilita este cálculo. Assim, aplicando $\ln$ obtém-se a função de log-verossimilhança, dada por:

$$\ell_i(\theta) = \ln(L_i(\theta, \mathbf{x})) \quad (2)$$

- No caso de variáveis aleatórias normalmente e identicamente distribuídas, a função de log-verossimilhança pode ser escrita como:

$$\ell_i(\theta) = \sum_{i=1}^N \log(f(\theta, x_i)) \quad (3)$$

Estimação com uma distribuição de Bernoulli

- Considere uma distribuição do Tipo Bernoulli. Nesse caso, a função de densidade de probabilidade é:

$$f(x_i, \theta) = \theta^x (1 - \theta)^{1-x}, 0 < \theta < 1 \quad (4)$$

- Em que θ é o parâmetro escalar que representa a probabilidade de sucesso. Nesse caso, os valores da amostra serão uma sequência de zeros e uns, denotando o fracasso e o sucesso de cada experimento, respectivamente.
- A função de log-verossimilhança será:

$$\ell_i(\theta) = \left(\sum_{i=1}^N x_i \right) \ln(\theta) + \left(N - \sum_{i=1}^N x_i \right) \ln(1 - \theta) \quad (5)$$

- Derivando em relação a θ e igualando a zero, tem-se:

$$\frac{\partial \ell_i(\theta)}{\partial \theta} = \left(\sum_{i=1}^N x_i \right) \frac{1}{\theta} - \left(N - \sum_{i=1}^N x_i \right) \frac{1}{1 - \theta} = 0 \quad (6)$$

Estimação com uma distribuição de Bernoulli

- Resolvendo a equação anterior para θ , obtém-se:

$$\begin{aligned}\left(\sum_{i=1}^N x_i\right) \frac{1}{\theta} &= \left(\frac{1}{1-\theta}\right) \left(N - \sum_{i=1}^N x_i\right) \\ \sum_{i=1}^N x_i &= \frac{\theta}{1-\theta} \left(N - \sum_{i=1}^N x_i\right) \\ (1-\theta) \left(\sum_{i=1}^N x_i\right) &= \theta N - \theta \left(\sum_{i=1}^N x_i\right) \\ \sum_{i=1}^N x_i &= \theta N \\ \hat{\theta} &= N^{-1} \left(\sum_{i=1}^N x_i\right)\end{aligned}\tag{7}$$

Estimação com distribuição exponencial

- Considere T como sendo o tempo de uso de uma peça até a sua falha. Suponha também que T tem uma distribuição esponencial com fdp dada por:

$$f(t) = \beta e^{-\beta t} \quad (8)$$

- Suponha que n dessas peças estejam ensaiadas, fornecendo as durações até a falha $T_1, T_2, \dots, T_n$.
- A função de verossimilhança desta amostra será:

$$L(T_1, T_2, \dots, T_n) = \beta^n \exp\left(-\beta \sum_{i=1}^n T_i\right) \quad (9)$$

- A função de log-verossimilhança desta amostra será:

$$\ell(T_1, T_2, \dots, T_n) = n \ln(\beta) - \beta \sum_{i=1}^n T_i \quad (10)$$

Estimação com distribuição exponencial

- Derivando em relação a β obtém-se:

$$\begin{aligned}\frac{\partial \ell(T_1, T_2, \dots, T_n)}{\partial \beta} &= \frac{n}{\beta} - \sum_{i=1}^n T_i = 0 \\ \hat{\beta} &= \frac{n}{\sum_{i=1}^n T_i} \\ \hat{\beta} &= \frac{1}{\bar{T}}\end{aligned}\tag{11}$$

- Em que $\bar{T} = \frac{\sum_{i=1}^n T_i}{n}$.

Estimação com distribuição normal

- Considere uma amostra com distribuição normal. A função de verossimilhança é:

$$L(x, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \quad (12)$$

- Aplicando Logaritmo para encontrar a função de log-verossimilhança:

$$\ell(x, \mu) = -\ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{(x_i - \mu)^2}{2\sigma^2} \quad (13)$$

- Pelas propriedades do logaritmo, $\ln(\sigma) = \frac{1}{2}\ln(\sigma^2)$. com isso, a função de log-verossimilhança pode ser escrita como:

$$\ell(x, \mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 \quad (14)$$

Estimação com distribuição normal

- Derivando em relação a μ :

$$\begin{aligned}\frac{\partial \ell(x, \mu)}{\partial \mu} &= \frac{-1}{2\sigma^2} 2 \sum_{i=1}^N (x_i - \mu)(-1) = 0 \\ \frac{1}{2\sigma^2} (x_1 - \mu + x_2 - \mu + \dots + x_N - \mu) &= 0 \\ \sum_{i=1}^N (x_i) - N\mu &= 0 \\ \hat{\mu} &= \frac{\sum_{i=1}^N (x_i)}{N}\end{aligned}\tag{15}$$

- O que demonstra que o valor ótimo de μ independe de σ .

Estimação de mínimos quadrados ordinários com erros normalmente distribuídos

- Considere uma equação de uma regressão linear múltipla.

$$\begin{aligned} y &= X\beta + u \\ u | X &\sim N(0, \sigma^2 I_T) \end{aligned} \quad (16)$$

- Como os erros são normalmente distribuídos, o processo de minimização dos resíduos quadráticos pode ser construído com base em uma função de distribuição de probabilidade normal, a qual é escrita como:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (17)$$

- Como μ é a média e $x - \mu$ é equivalente ao erro da estimativa, e sabendo também que os resíduos quadráticos são:

$$e'e = (y - X\beta)'(y - X\beta) \quad (18)$$

Estimação de mínimos quadrados ordinários com erros normalmente distribuídos

- A função de verossimilhança pode ser escrita como:

$$f(x, \beta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}} \quad (19)$$

- A função de log-verossimilhança é:

$$\ell(x, \beta) = -\ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{(y-X\beta)'(y-X\beta)}{2\sigma^2} \quad (20)$$

- Pelas propriedades do logaritmo, $\ln(\sigma) = \frac{1}{2}\ln(\sigma^2)$. com isso, a função de log-verossimilhança pode ser escrita como:

$$\ell(x, \beta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y-X\beta)'(y-X\beta)}{2\sigma^2} \quad (21)$$

- Resolvendo o produto $(y-X\beta)'(y-X\beta)$, obtêm-se:

$$\ell(x, \beta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{[y'y - 2\beta'X'y + \beta'X'X\beta]}{2\sigma^2} \quad (22)$$

Estimação de mínimos quadrados ordinários com erros normalmente distribuídos

- Derivando em relação a β e aplicando a igualdade em zero, obtém-se:

$$\begin{aligned}\frac{1}{2\sigma^2} [-2X'y + 2X'X\beta] &= 0 \\ X'y &= X'X\beta \\ \beta &= (X'X)^{-1} (X'y)\end{aligned}\tag{23}$$

- O que demonstra que o princípio da minimização dos quadrados dos resíduos também pode ser efetuada pelo processo de máxima verossimilhança, produzindo os mesmos parâmetros desde que os erros sejam normalmente distribuídos.

Estimação do modelo Logit

- Considere uma amostra aleatória com uma variável binária Y em que $Y = 1$ representa o sucesso e $Y = 0$ representa o fracasso.
- Agora imagine que a probabilidade de sucesso está condicionada a um conjunto X de variáveis.
- Considere também que se deseja escrever a probabilidade de sucesso como uma função linear das variáveis condicionais Nesse caso:

$$P(Y = 1 | X) = \sum_{i=1}^N \beta_i x_i \quad (24)$$

- Com $x_1 = 1$. Considere, no entanto, que a probabilidade de sucesso assume a forma de uma distribuição logística, ou seja:

$$P(Y = 1 | X = x) = \frac{1}{1 + e^{\beta_i x_i}} \quad (25)$$

- A probabilidade de fracasso pode ser escrita como:

$$P(Y = 0 | X = x) = 1 - p \Rightarrow 1 - \frac{1}{1 + e^{\beta_i x_i}} \Rightarrow \frac{e^{\beta_i x_i}}{1 + e^{\beta_i x_i}} \quad (26)$$

Estimação do modelo Logit

- Em que β é um vetor de parâmetros desconhecidos a ser estimado.
- O próximo passo é generalizar a probabilidade na forma de uma função do tipo Bernoulli:

$$\begin{aligned} P(Y = y \mid X = x) &= \left(\frac{1}{1 + e^{\beta_i x_i}} \right)^{y_i} \left[1 - \frac{1}{1 + e^{\beta_i x_i}} \right]^{(1-y_i)} \\ P(Y = y \mid X = x) &= \left(\frac{1}{1 + e^{\beta_i x_i}} \right)^{y_i} \left[\frac{e^{\beta_i x_i}}{1 + e^{\beta_i x_i}} \right]^{(1-y_i)} \end{aligned} \quad (27)$$

- Seguindo essa especificação, é possível escrever a função de verossimilhança como:

$$P(Y = y \mid X = x) = \prod_{i=1}^N \left(\frac{1}{1 + e^{\beta_i x_i}} \right)^{y_i} \left[\frac{e^{\beta_i x_i}}{1 + e^{\beta_i x_i}} \right]^{(1-y_i)} \quad (28)$$

- O uso do produtório indica que o modelo considera a suposição de que a amostra aleatória é IID.

Estimação do modelo Logit

- O próximo passo é aplicar o logaritmo para obter a função de log-verossimilhança.

$$\ell(\beta) = \sum_{i=1}^N \left[y_i \ln \left(\frac{1}{1 + e^{\beta_i x_i}} \right) + (1 - y_i) \ln \left[\frac{e^{\beta_i x_i}}{1 + e^{\beta_i x_i}} \right] \right] \quad (29)$$

- Derivando em relação a θ , obtém-se:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^N \left[-\frac{\beta y_i e^{\beta x}}{1 + e^{\beta x}} + \frac{\beta(1 - y_i)}{1 + e^{\beta x}} \right] = 0 \quad (30)$$

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

$$\sum_{i=1}^N [y_i F(-x\beta) - (1 - y_i) F(x\beta)] = 0 \quad (31)$$

- O valor de β pode ser obtido por meio do método de Newton–Raphson para encontrar um vetor de valores que torne a equação anterior igual a zero.

Estimação do modelo Probit

- O modelo probit segue a mesma lógica do modelo logit, alterando-se apenas no que diz respeito à função densidade de probabilidade, e consequentemente, à função de verossimilhança.
- Enquanto o modelo logit considera que a probabilidade assume uma distribuição do tipo logística, o modelo probit considera que o vetor de probabilidades estimadas segue uma distribuição normal padrão.
- Para representar, considere η como sendo o preditor linear que determina a especificação da probabilidade estimada.

$$\eta_i = \eta_i(X_i, \beta) = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} \quad (32)$$

- No modelo probit, a probabilidade de sucesso pode ser escrita como:

$$P(Y = 1 | X) = \Phi(\eta) = \int_{-\infty}^{\eta} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \quad (33)$$

- Em que $\frac{e^{-z^2/2}}{\sqrt{2\pi}}$ é a função de densidade de probabilidade normal padrão e $z \in \mathbb{R}$.

Estimação do modelo Probit

- Sabendo que a probabilidade de fracasso é $1 - P(Y = 1 | X)$, o próximo passo é generalizar a probabilidade na forma de uma função do tipo Bernoulli.

$$P(Y = y | X = x) = \Phi(\eta_i)^{Y_i} (1 - \Phi(\eta_i))^{1-Y_i} \quad (34)$$

- A função de verossimilhança pode ser escrita como:

$$L(\beta | (X_i, Y_i)) = \prod_{i=1}^n \Phi(\eta_i)^{Y_i} (1 - \Phi(\eta_i))^{1-Y_i} \quad (35)$$

- A função de log-verossimilhança pode ser escrita como:

$$\ell(\beta) = \sum_{i=1}^n Y_i \log(\Phi(\eta_i)) + (1 - Y_i) \log(1 - \Phi(\eta_i)) \quad (36)$$

Estimação do modelo Probit

- Derivando a função de log-verossimilhança em relação à β , tem-se:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^n X_i \cdot \phi(\eta_i) \left(\frac{Y_i}{\Phi(\eta_i)} - \frac{1 - Y_i}{1 - \Phi(\eta_i)} \right) \\ &= \sum_{i=1}^n X_i \cdot \phi(\eta_i) \left(\frac{Y_i}{\Phi(\eta_i)(1 - \Phi(\eta_i))} - \frac{1}{1 - \Phi(\eta_i)} \right)\end{aligned}\quad (37)$$

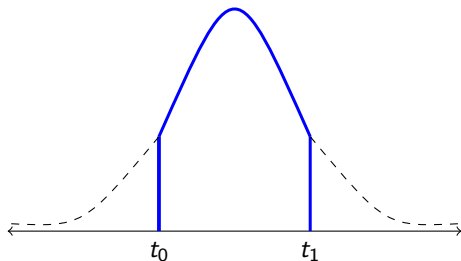
- Lembre-se que:

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \quad \text{e que} \quad \Phi(\eta) = \int_{-\infty}^{\beta_0 + \sum_{j=1}^p \beta_j X_{ij}} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \quad (38)$$

- Assim como no modelo logit, o método de Newton-Raphson pode ser usado para encontrar o valor de β que torne a Equação 37 igual a zero.

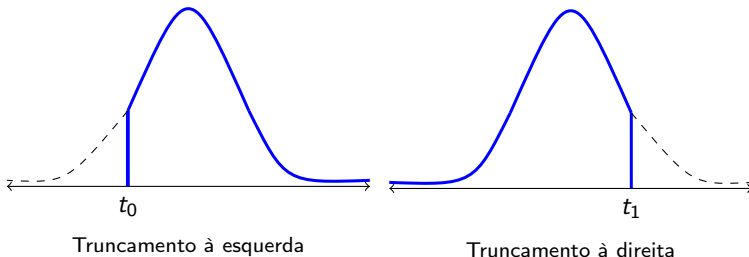
Estimação do modelo Tobit

- A lógica do modelo tobit é semelhante à lógica dos modelos logit e probit. No entanto, nesse caso considera-se uma função densidade de probabilidade normal e adiciona-se o conceito de censura e truncamento.
- Imagine que uma variável aleatória Y segue uma distribuição normal e é **truncada** no intervalo (t_0, t_1) . Nesse caso, tem-se:
- Nesse caso, Y não existe fora do intervalo (t_0, t_1) .
- Ex: Proporções; Nota do Enem; Índice de Gini.



Estimação do modelo Tobit

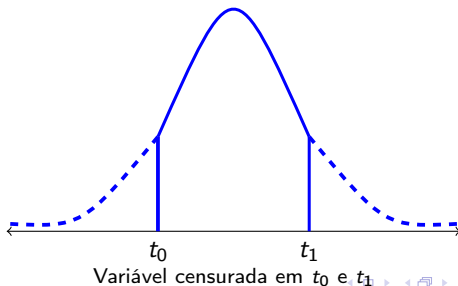
- O truncamento demonstrado anteriormente ocorre tanto na parte inferior quanto na parte superior do limite.
- Também existe o **truncamento à esquerda** (ou para baixo) que ocorre quando a variável aleatória não possui valores menores do que o limite inferior do truncamento mas o limite superior é infinito.
- E o **truncamento à direita** (ou para cima), que ocorre quando a variável aleatória não possui valores maiores do que o limite superior do truncamento, mas o limite inferior é infinito.



- Exemplo de variáveis truncadas à esquerda: Idade (Truncada em zero), Salário (Truncada em zero), Mão de obra (Truncada em zero), tecnologia (Truncada em zero).
- Exemplo de variáveis truncadas à direita:

Estimação do modelo Tobit

- A censura é bastante semelhante ao truncamento, porém, na censura os valores existem fora dos limites considerados, no entanto, não são observados.
- Imagine uma variável que representa a remuneração do trabalho dos indivíduos que ganham entre 1 e 2 salários mínimos. Existem remunerações maiores que 2 salários mínimos assim como existem remunerações menores que 1 salário mínimo. Porém, esses valores não são considerados. Neste caso, a remuneração estaria sendo censurada no intervalo de 1 e 2 salários mínimos.



Estimação do modelo Tobit

- Para demonstrar os procedimentos de cálculo do modelo tobit, considere uma variável aleatória y com truncamento ou censura em um limiar Γ à esquerda. Nesse caso, defina d como uma condição binária de tal forma que:

$$d = \begin{cases} 1 & \text{se } y > \Gamma \\ 0 & \text{se } y \leq \Gamma \end{cases} \quad (39)$$

- Agora considere que a variável y pode ser escrita como uma função linear dos seus regressores:

$$y = \sum_{i=1}^N \beta_i X_i \quad (40)$$

- Agora assuma que a probabilidade de ocorrência de $y > \Gamma$ pode ser representada por uma distribuição normal de tal forma que:

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} \quad (41)$$

- Reescreva 41 de tal modo que:

$$\begin{aligned} p(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \sum_{i=1}^N \beta_i X_i)^2}{2\sigma^2}} \\ &= \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{y_i - \sum_{i=1}^N \beta_i X_i}{\sigma} \right)^2} \end{aligned} \quad (42)$$

- Para $y \leq \Gamma$, tem-se:

$$\begin{aligned} p(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(-\sum_{i=1}^N \beta_i X_i)^2}{2\sigma^2}} \\ &= \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{-\sum_{i=1}^N \beta_i X_i}{\sigma} \right)^2} \end{aligned} \quad (43)$$

Estimação do modelo Tobit

- A função de verossimilhança pode ser escrita como:

$$L = \prod_{i=1}^N \left[\frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{y_i - \sum_{i=1}^N \beta_i X_i}{\sigma} \right)^2} \right]^d \left[1 - \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{-\sum_{i=1}^N \beta_i X_i}{\sigma} \right)^2} \right]^{1-d}$$

- A função de log-verossimilhança será:

$$\ell(\beta, \sigma^2) = \sum_{i=1}^N \left\{ d_i \left(-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left(y_i - \sum_{i=1}^N \beta_i X_i \right)^2 \right) + \right. \\ \left. (1 - d_i) \left(-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left(-\sum_{i=1}^N \beta_i X_i \right)^2 \right) \right\}$$

Estimação do modelo Tobit

- Derivando em relação a β :

$$\frac{\partial \ln \ell}{\partial \beta} = \sum_{i=1}^N \frac{1}{\sigma^2} \left(d_i \left(y_i - \sum_{i=1}^N \beta_i X_i \right) + (1 - d_i) \left(- \sum_{i=1}^N \beta_i X_i \right) \right) (-x_i) = 0$$

- Novamente, o método de Newton-Raphson pode ser usado para encontrar o valor de β que torna a Equação anterior igual a zero.
- No caso da censura ou truncamento à direita, o procedimento é o mesmo, apenas o valor de d é alterado, passando a ser definido como:

$$d = \begin{cases} 1 & \text{se } y < L \\ 0 & \text{se } y \geq L \end{cases} \quad (44)$$

- Similarmente, quando existe truncamento ou censura à direita e à esquerda, d passa a ser escrito como:

$$d = \begin{cases} 1 & \text{se } L_0 < y < L_1 \\ 0 & \text{se } y \leq L_0 \text{ ou } y \geq L_1 \end{cases} \quad (45)$$