Trabalho de Inferência

Modelo Espacial AutoRegressivo (SAR)

Rafael Morciani/GRR:20160217

2017-12-01

Resumo

O modelo de regressão espacial SAR é um modelo estatístico que tem como objetivo medir a dependência espacial de uma variavél explicativa com uma variável a ser explicada e/ou a relação entre áreas com a mesma variável, como por exemplo: explicar a renda em relação ao grau de escolaridade; verificar se a renda em uma determinada área é relacionada com a renda de áreas vizinhas. O modelo é utilizado em espaços geográficos, não levando em consideração relevos (ele atua no plano), então ele mede a dependência e/ou relação no espaços.

Introdução

O modelo consiste em mensurar o grau de dependência entre a variável dependente das variáveis independentes e/ou a relação de uma única variável em diferentes regiões. O modelo é utilizado em políticas públicas (saúde, econômia, escolaridade, etc), como por exemplo: Explicar o nível de escolaridade de uma determinada região pelo salário médio dos vizinhos desta região. O modelo é muito utilizado para planejamento, como por exemplo: prever a econômia de um local que ainda não possuem moradores com base na econômia das áreas vizinhas.

Os dados utilizados na implementação do modelo serão obtidos a partir de simulações no R, após os dados serem simulados podemos montar as funções de log e verossimilhança. As derivadas destas funções são muito complexas, então a função escore não será cálculada, ao invés utilizaremos a função "optim" do R para maximizar a log-verossimilhança em função dos parâmetros μ, ρ, σ^2 *e* $\beta's$ obtendo assim os estimadores de máxima verissimilhança do modelo ($\hat{\mu}, \hat{\rho}, \hat{\sigma^2}$ *e* $\hat{\beta's}$).

O modelo tem como objetivo medir a dependência/relação espacial, ou seja os valores de ρ, caso ρ seja muito próximo de zero, nosso modelo de regressão se torna ou se aproxima de um modelo de regressão linear.

Modelo

O modelo SAR Espacial consiste pelas equações:

$\mathrm{Y} = \rho\, W\, \mathrm{Y} + \mathrm{X}\, \beta + \epsilon,\ \epsilon \sim N(0; \sigma^2 I)$

Onde:

Y:Variável dependente;

ρ:Parâmetro espacial responsável por mensurar o grau de dependência espacial;

W:Matriz de vizinhança;

X: Vetor das variáveis independentes;

β:Vetor dos coeficientes de regressão;

ϵ: Erro aleatório;

σ^2:Variância do modelo;

I: Matriz identidade.

Suporte da distribuição: $\mathrm{Y} \in \mathbb{R}$

Espaço paramétrico: $\rho \in [-1; 1],\ \sigma^2 \in \mathbb{R}_+^*,\ \beta \in \mathbb{R},\ \mu \in \mathbb{R}$

Inferência

Para a devida inferência do modelo SAR espacial precisamos encontrar os estimadores de cada um dos parâmetros (μ, ρ , σ^2 e $\beta's$).

A estimação de ρ será realizada pela máxima verossimilhança da distribuição normal multivariada:

Z $\sim N_m(\mu; \Sigma)$

$\Sigma = (I - \rho W)^{-1}\sigma^2 I(I - \rho W')^{-1}$

Assim ficamos com a seguinte expressão:

$$f(y_i; \rho; \mu; \sigma^2) = (2\pi)^{\frac{-p}{2}} \, |\Sigma|^{\frac{-1}{2}} \, e^{\frac{-1}{2}(y_i-\mu)\Sigma^{-1}(y_i-\mu)'}$$

Onde p = Total de observações

Como cada área possui somente uma observação (com vários dados) então n=1. E nossa função de verossimilhança fica igual a função distribuição.

Para medir a compatibilidade do parâmetro com a amostra obtida, utilizaremos o método de máxima verossimilhança.

Verossimilhança

$$L(\rho; \mu; \sigma^2; y_i) = (2\pi)^{\frac{-p}{2}} \, |\Sigma|^{\frac{-1}{2}} \, e^{\frac{-1}{2}(y_i-\mu)\Sigma^{-1}(y_i-\mu)'}$$

log-verossimilhança

$$l(\rho; \mu; \sigma^2; y_i) = -\frac{1}{2}\left[p\log(2\pi) + log(|\Sigma|) + (Y - \mu)\Sigma^{-1}(Y - \mu)'\right]$$

Maximizando a função

Para maximizar a log-verossimilhança usaremos a função "optim", com ela conseguiremos os estimadores de todos os parâmetros do modelo ($\hat{\mu}, \hat{\rho}, \hat{\sigma^2}$ e $\hat{\beta's}$).

Após calculado os estimadores, verificamos que os mesmos são consistentes, não viciados e eficientes.

Com os valores dos estimadores dos parâmetros, iremos criar um intervalo com 95% de confiança de conter o verdadeiro valor do parâmetro, realizaremos também um teste de hipótese (Wald) para saber o quão distante os estimadores estão dos valores observados.

Teste Wald: $Z_n = \frac{\hat{\theta}-\theta}{\sqrt{V[\hat{\theta}]}} \sim N(0; 1)$

Se $Z_n > 1,96$ ou $Z_n < -1,96$ então, rejeita-se a hipótese de que $\rho = 0$

Implementação Computacional

1. **Simulando os dados**

```r
# Criando a estrutura espacial
# Lista de vizinhos
nb <- cell2nb(nrow = 10, ncol = 10)

# Matriz de vizinhanca
matriz = cell2nb(nrow=10,ncol=10,type="rook")
W = nb2mat(matriz)

#Demais variaveis
Sigma <- function(rho, sigma2, W) {
  I <- diag(rep(1,ncol(W)))
  B <- rho*W
  p1 <- solve( (I - B) )
  p2 <- solve( (I - t(B)) )
  out <- sigma2*(p1%*%p2)
  return(out)
}

Sigma2 <- as.matrix(Sigma(rho = 0.5, sigma2 = 1, W = W))

x1 <- seq(-1,1, l = 100)

mu <- 2 + 3*x1

Y <- as.numeric(rmvnorm(1, mean = mu, sigma = Sigma2))
```

2. **Função de Verossimilhança e Log-Verossimilhança**

```
#Verossimilhanca
L <- function(Sigma,Y) {
  p <- length(Y)
  out <- prod( ((2*pi)^(-p/2)) * (1/determinant(Sigma,logarithm=T)) *
                 exp((-1/2)*Y%*%solve(Sigma)%*%t(Y)) )
  return(out)
  }
```

Para melhor manipulação da função, passamos o logaritmo em ambos os lados da função, ficando com a função de log-verossimilhança.

```
#Log-verossimilhanca
lv <- function(par, W, Y, x1) {
  beta0 <- par[1]
  beta1 <- par[2]
  rho <- par[3]
  sigma2 <- par[4]
  p <- length(Y)
  Sigma2 <- as.matrix(Sigma(rho = rho, sigma2 = sigma2, W = W))
  mu <- beta0 + beta1*x1
  out <- dmvnorm(x = Y, mean = mu, sigma = Sigma2, log = TRUE)
  return(as.numeric(out))
}
```

3. **Maximizando a função Log-Verossimilhança pela função "optim"**

Para utilização desta função, precisou definir o espaço paramétrico de cada estimados, para isso foi utilizado o método "L-BFGS-B".

```
est <- optim(par = c(0,0, 0, 1), fn = lv, W = W, Y = Y, x1 = x1,
             method = "L-BFGS-B",
             lower = c(-Inf, -Inf, -0.99, 0.0001), upper = c(Inf, Inf, 0.99, Inf),
             control = list(fnscale = -1), hessian = T)
```

4. **Estimativas obtidas**

```
Beta0_hat <- est$par[1]
Beta0_hat
```

```
## [1] 2.005247
```

```
Beta1_hat <- est$par[2]
Beta1_hat
```

```
## [1] 3.251509
```

```
rho_hat <- est$par[3]
rho_hat
```

```
## [1] 0.3544561
```

```
sigma2_hat <- est$par[4]
sigma2_hat
```

```
## [1] 0.9230896
```

5. **Matriz de informação observada e as variâncias dos estimadores**

A matriz de informação observada possiu na diagonal principal as variâncias de cada estimador e os demais valores são as covariâncias dos estimadores.

```
#Matriz de informação observada
I_o <- (-est$hessian)
I_o
```

```
##               [,1]          [,2]        [,3]          [,4]
## [1,]  4.514480e+01 -7.105427e-09  0.50547543 -8.322587e-05
## [2,] -7.105427e-09  1.604223e+01 -0.01084038 -2.776801e-05
## [3,]  5.054754e-01 -1.084038e-02 69.01823112  1.175495e+01
## [4,] -8.322587e-05 -2.776801e-05 11.75494508  5.868004e+01
```

```
V <- sqrt(diag(solve(I_o)))
V
```

```
## [1] 0.1488383 0.2496708 0.1224826 0.1328292
```

6. **Teste de hipótese e intervalo com 95% de confiança**

Teste Wald: $H_0 : \rho = 0,\ H_1 : \rho \neq 0$

```
rho_0 <- 0
Zn <- (rho_hat - rho_0)/(V[3])
Zn
```

```
## [1] 2.893929
```

Intervalo com 95% de confiança

```
inf <- rho_hat - (1.96*V[3])
sup <- rho_hat + (1.96*V[3])
IC <- c(inf,sup)
IC
```

```
## [1] 0.1143901 0.5945221
```

Discussão

- Valores definidos para os estimadores:

$\rho = 0.5$

$\sigma^2 = 1$

$\beta_0 = 2$

$\beta_1 = 3$

- Estimativas obtidas:

$\hat{\rho} = 0.3544561$

$\hat{\sigma^2} = 0.9230896$

$\hat{\beta_0} = 2.0052472$

$\hat{\beta_1} = 3.2515094$

- Teste Wald

$H_0 : \rho = 0$, $H_1 : \rho \neq 0$

$Z_n = 2.8939293$

Rejeita-se H_0

- Intervalo com 95% de confiança para $\hat{\rho}$

$IC_{95\%}$: [0.1143901, 0.5945221]

A inferência realizada no modelo SAR espacial utilizando o método de máxima verossimilhança a partir de simulações realizadas no R, se mostrou muito eficiente, conseguindo medir com precisão se existe ou não dependência espacial. Mesmo não realizando todos os cálculos analiticamente conseguimos boas estimativas.

O grande desafio foi trabalhar com um modelo pouco conhecido e na simulação dos dados, mas uma vez que os dados foram gerados, e com a distribuição conhecida, foi possível realizar a devida inferência. A função "optim" também se mostrou efetiva no cálculo da maximização da função, retornando todas as estimativas necessárias para a realização do teste de hipótese e criação do intervalo de confiança.