Regressão segmentada com nodos desconhecidos em modelos lineares generalizados

Fernando Lucambio Pérez
Departamento de Estatística
Universidade Federal do Paraná

Abril de 2004

Introdução

Em analises de regressão a variável resposta Y é modelada como uma função das variáveis explicativas $x_1, x_2, \cdots$. As vezes acontece que a relação entre a variável resposta e alguma das variáveis explicativas mostra alguns valores onde o efeito na resposta muda abruptamente.

Essse valores são conhecidos como nodos ou pontos de transição e modelos de regressão com nodos são chamados de regressão segmentada.

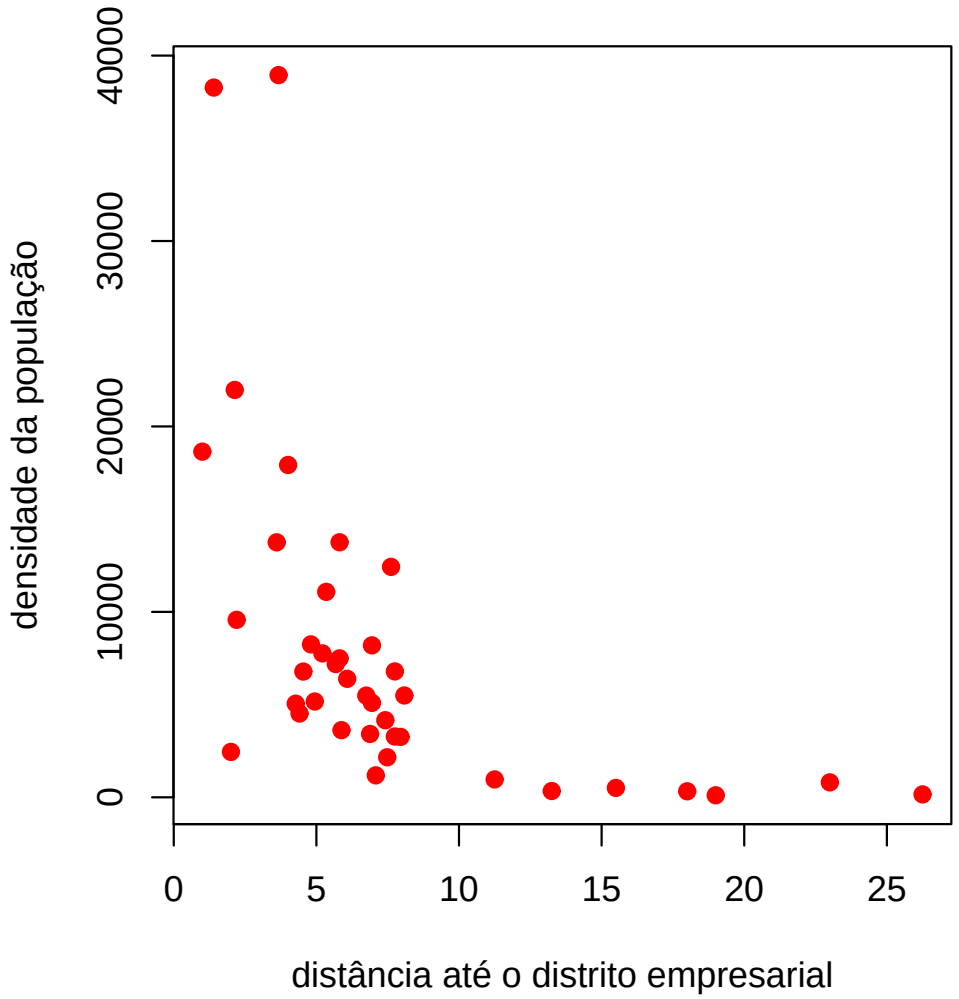
Quando os nodos são desconhecidos precissamos estimálos e os métodos inferencias clássicos não são aplicáveis.

Mostraremos aqui métodos alternativos de estimação tanto dos parâmetros de regressão quanto dos nodos com ênface nas aplicações. Mostraremos os problema de regressão segmentada através do seguinte exemplo.

Dados sobre densidade populacional em diferentes setores censitários da área de Baltimore em 1970.

Variáveis: densidade da população no setor censitário e distância entre o setor censitário e o distrito empresarial

Publicado em K. Lahiri and R. Numrich, ''An Econometric Study os the Dynamics of Urban Spatial Structure'', Journal of Urban Economics, 1983, pp. 55-79.


densidade da população
0
10000
20000
30000
40000
0
5
10
15
20
25
distância até o distrito empresarial

Modelo de regressão segmentada

Consideremos o seguinte modelo de regressão,

$$\mathrm{E}\{Y\} = \begin{cases} g(\alpha_1 + \beta_1 x) & \text{se } x \leq \tau_1, \\ g(\alpha_2 + \beta_2 x) & \text{se } \tau_1 < x \leq \tau_2, \\ \vdots & \vdots \\ g(\alpha_K + \beta_K x) & \text{se } \tau_{K-1} < x, \end{cases}$$

onde $f(y;\theta,\phi) = \exp\{\phi(y\theta + b(\theta) + c(y)) + d(y,\phi)\}$.

Os pontos finais dos intervalos denotam os nodos e dado que são desconhecidos, tem que ser estimados. Por razões teóricas e práticas supõe-se que os nodos estão entre o menor e o maior valor de x.

Assumiremos que o modelo de regressão segmentado é contínuo, ou seja,

$$\alpha_i + \beta_i \tau_i = \alpha_{i+1} + \beta_{i+1} \tau_i,$$

para todo $i = 1, \cdots, K-1$ sendo K o número de nodos.

O modelo pode ser escrito segundo uma outra parametrização da forma,

$$\mathrm{E}\{Y\} = g(\alpha + \beta_1 x + \sum_{i=2}^{K} \beta_i (x - \tau_{i-1})_+),$$

onde

$$t_+ = \begin{cases} t & \text{se } t \geq 0 \\ 0 & \text{se } t < 0. \end{cases}$$

Exemplo: densidade da população em relação à distância ao distrito empresarial central

```
glm(formula = y ~ x, family = Gamma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9208  -0.6045  -0.1307   0.2414   1.7931

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.832e-06  1.630e-05  -0.603     0.55
x            3.057e-05  5.394e-06   5.667 1.78e-06 ***

(Dispersion parameter for Gamma family taken to be 0.6457577)

    Null deviance: 51.770  on 38  degrees of freedom
Residual deviance: 26.565  on 37  degrees of freedom
AIC: 755.48
```

Regressão segmentada

```
segmented.glm(obj = adj1, Z = x, psi = 7)

Estimated Break-Point(s):
    Est St.Err
7.6490 0.4321

Meaningful coefficients of the linear terms:
                Estimate   Std. Error   t value
(Intercept) 1.239231e-05 2.012296e-05 0.6158296
x           2.077250e-05 5.684837e-06 3.6540196
U.x         1.942193e-04 6.493095e-05 2.9911670

(Dispersion parameter for Gamma family taken to be 0.498328)

    Null deviance: 51.770  on 38  degrees of freedom
Residual deviance: 15.645  on 35  degrees of freedom
AIC: 737.06
```

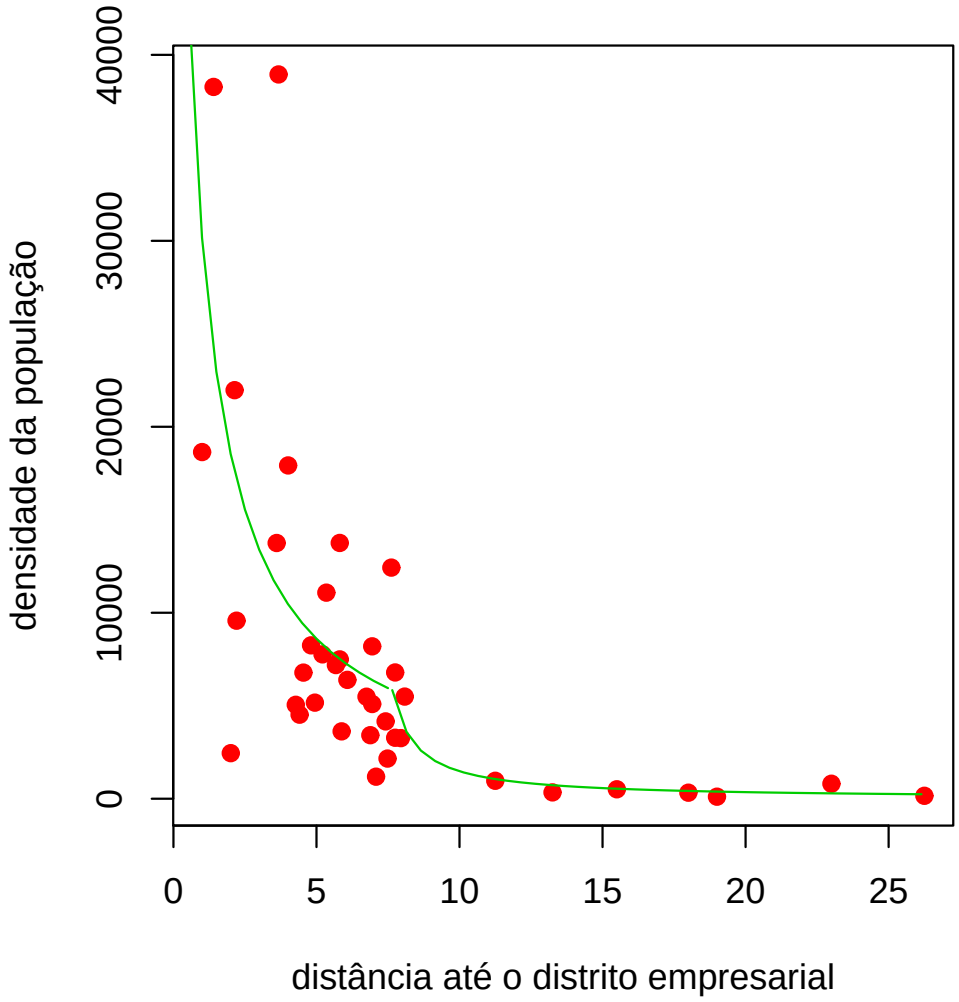
Também temos a disposição a função **slope.segmented** a qual nos permite obter as estimativas do coeficientes β.

```
                Est.      St.Err.  t value    CI(95%).l    CI(95%).u
slope1 2.077250e-05 5.684837e-06 3.654020 9.630429e-06 3.191458e-05
slope2 2.149918e-04 6.468162e-05 3.323848 8.821819e-05 3.417655e-04
```

As retas de regressão estimadas são,

$$\widehat{\mathrm{E}\{Y\}} = 1/(1.239231e-05 + 2.077250e-05 * x) \quad \text{se} \quad x < 7.649,$$
$$\widehat{\mathrm{E}\{Y\}} = 1/(-0.001473191 + 2.149918e-04 * x) \quad \text{se} \quad x \geq 7.649,$$

para as quais utilizamos as restrições de continuidade.


densidade da população
0
10000
20000
30000
40000
0
5
10
15
20
25
distância até o distrito empresarial

Referências bibliográficas

- Hinkley, D. (1971). Inference in two-phase regression. JASA.

- Muggeo, V. (2003). Estimating regression models with unknown break-points. Statistics in Medicine.

- Seber, G. e Wild, C. (1989). Nonlinear regression. John Wiley & Sons.