Análise e planejamento estatístico de rotas

Introdução

A análise de rotas tem grande potencial para identificar pontos de vulnerabilidade em um rede de mobilidade urbana. Com o advento dos dados de torres de telefones celulares, dispomos hoje de dados ricos sobre padrões de movimento urbano através de dados de origem-destido (OD). Nesta lição prática, vamos utilizar análise estatística exploratória e modelos estatísticos para analisar rotas de mobilidade urbana em uma importante região metropolitana do Brasil.

Análise de rotas : dados origem-destino

Nesta seção vamos analisar dados de origem-destino apresentados em Chaves et al. (2023). Em essência, os dados descrevem o número de indivíduos que viajam entre duas localidades em um determinado período (no nosso caso, um dia). Além disso, temos a distância entre duas localidades medida de três formas diferentes: distância euclidiana, distância a pé (*walking*) e distância de carro (*driving*). Temos também as populações de cada área – ver modelo gravitacional abaixo.

O modelo gravitacional

Mais detalhes em Santos et al. (2019).

Seja F_{ij} o fluxo entre duas áreas (nós) i e j, e sejam P_i e P_j suas respectivas populações. Por fim, seja d_{ij} a distância entre os nós.

O modelo gravitacional diz que

$$F_{ij} = \alpha \cdot \frac{P_i \cdot P_j}{(d_{ij})^\gamma}, \tag{1}$$

onde α e γ são parâmetros livres a serem estimados. Nos exercícios propostos abaixo, você vai ajustar este modelo aos dados através de uma transformação que torna o problema de ajuste um problema de regressão.

Modelos lineares generalizados (GLM)

Sejam $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ e $\boldsymbol{X}$ o vetor de variáveis dependentes e a matriz $(n \times P)$ de desenho, respectivamente. Defina $\mu_i(\boldsymbol{X}) = \mu_i := E[Y_i \mid \boldsymbol{X}]$ como a média condicional de cada Y_i. Em um GLM, escrevemos

$$g(\mu_i) = \boldsymbol{X}\boldsymbol{\beta}, \tag{2}$$

onde $g()$ é a chamada *função de ligação*, e é uma função monotônica e diferenciável. Por exemplo, quando $Y_i \in \{0, 1\}$, isto é, quando temos dados binários, podemos usar $g(\theta) = \log(\theta/(1-\theta))$ ou $g(\theta) = \Phi^{-1}(\theta)$, onde Φ^{-1} é a CDF inversa de uma distribuição normal padrão. Além disso quando temos dados de contagem ($Y_i \in \{0, 1, \ldots\}$) distribuídos Poisson – por exemplo –, é de costume usar $g(\theta) = \log(\theta)$.

Análise dirigida

1. Carregue os dados e investigue sua forma e dimensões.
2. Calcule as porcentagens de zeros em cada variável.
3. Filtre as entradas do banco correspodentes às entradas fora da diagonal na matriz origem-destino.
4. Investigue a correlação entre as variáveis de distância.
5. Sob o modelo gravitacional, mostre o que acontece quando consideramos a transformação $Z_i = \log(Y_i)$ e regredimos $\boldsymbol{Z}$ em $\boldsymbol{X}$.
6. Ajuste o modelo transformado aos dados discutidos acima. Alguma das distâncias dá resultados melhores? Como as estimativas dos parâmetros mudam?
7. Avalie a qualidade do ajuste dos modelos ajustados.
8. Desenhe uma estratégia para avaliar o poder preditivo dos modelos em novas amostras.

Referências

- Chaves, Júlio César, Moacyr AHB da Silva, Ricardo de Souza Alencar, Alexandre G. Evsukoff, and Vinícius da Fonseca Vieira. Human mobility and socioeconomic datasets of the Rio de Janeiro metropolitan area. Data in brief 51 (2023): 109695.
- Lima Santos, Leonardo Bacelar, Luiz Max Carvalho, Wilson Seron, Flávio C. Coelho, Elbert E. Macau, Marcos G. Quiles, and Antônio M. V. Monteiro. How do urban mobility (geo) graph's topological properties fill a map?. Applied Network Science 4 (2019): 1-14.
- Evans, M. J., & Rosenthal, J. S. (2018). Probability and Statistics - The Science of Uncertainty. CRC press. (Cap 10).

- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press. (Caps 3, 4 e 5)
- Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. Cambridge University Press. (Cap 15)