Sobredispersão em Modelos de Contagem

1 Motivação

O modelo de Poisson é amplamente usado para modelar dados de contagem, i.e., dados $Y \in \mathbb{N} \cup \{0\}$. O modelo de Poisson possui a famosa propriedade de *equidispersão*, ou seja de ter média e variância iguais:

$$Y \sim \text{Poisson}(\lambda) \Rightarrow \text{Var}(Y) = \mathbb{E}[Y] = \lambda.$$

Consequentemente, se nós acreditarmos que nossos dados seguem uma distribuição Poisson, esperamos que média e variância amostral sejam relativamente próximos. No entanto, é comum encontrar dados em que a dispersão é muito superior a média. Nesses casos um modelo GLM Poisson não tem flexibilidade suficiente para capturar esse fenômeno – note que o modelo implica a mesma estrutura para a média e a variância condicionais.

Este fenômeno é conhecido como **sobredispersão**, e se não for tratado de forma adequada, pode invalidar nossas inferências sob o modelo tradicional.

1.1 Questão 1 - Análise Exploratória

Para este exercício vamos considerar o conjunto de dados *RecreationDemand*[1] do pacote AER. Os dados são sobre o número de viagens recreativas de barco para o Lago Somerville, Texas, em 1980, com base em uma pesquisa aplicada a 2.000 proprietários de barcos de lazer registrados em 23 condados do leste do Texas.

Conduza uma análise exploratória sobre os dados. Verifique média e variância da variável *trips*. Investigue sua relação com outras variáveis disponíveis no banco e use esta análise para verificar a escolha ou não de covariáveis. Verifique a documentação disponível e descreva em detalhes as covariáveis escolhidas.

1.2 Questão 2 - Ajuste do Modelo Poisson

Ajuste um modelo de regressão de Poisson aos dados. Que métodos você usou e porquê? Os parâmetros se mostram significantes? Que métricas nos ajudam a entender o ajuste do modelo? Justifique cuidadosamente cada resposta.

1.3 Questão 3 - Um teste para sobredispersão

Cameron & Trivedi (1990) desenvolveram um teste de hipóteses para detectar a presença de sobredispersão. O teste parametriza a variância como

$$\text{Var}(Y) = \mu + \alpha \cdot \mu.$$

Assim definimos a hipótese nula e alternativa como:

$$\begin{aligned} H_0 : &\quad \alpha = 0 \\ H_1 : &\quad \alpha \neq 0 \end{aligned}$$

[1] `https://rdrr.io/cran/AER/man/RecreationDemand.html`

O teste é baseado no fato de que sobre o modelo Poisson

$$E[(Y-\mu)^2 - Y] = 0.$$

Fazemos o teste em 3 etapas:

- Etapa 1: Primeiro ajustamos um GLM Poisson e obtemos os valores de média ajustados $\hat{\mu}_i$.
- Etapa 2: Computamos a estatística de teste

$$Z_i = \frac{(Y_i - \hat{\mu}_i)^2 - Y_i}{\hat{\mu}\sqrt{2}}$$

- Etapa 3: Regredir Z_i em $\hat{\mu}_i$ sem intercepto e performar um teste-t no coeficiente da regressão. A significância do coeficiente indica a presença ou não de sobredispersão

Execute o teste nos dados e interprete os resultados.

1.4 Questão 4 - Modelos alternativos

Caso seja detectada a presença de sobredispersão, outros modelos podem ser utilizados para melhor ajustar os dados. Um exemplo seria trocar a Poisson por uma Binomial Negativa. Exiba uma parametrização conveniente da binomial negativa e explique porque ela seria conseguiria capturar a dispersão em excesso dos dados melhor que a Poisson. Em seguida, ajuste um GLM com a binominal negativa – no R, você pode usar a função `glm.nb()` do pacote **MASS**. Lembre-se de justificar a escolha de função de ligação.

1.5 Extra: Excesso de zeros

A sobredispersão pode ser causada por outros fenômenos em dados de contagem como excesso de zeros, trazendo a média próximo para o aglomerado de zeros. Uma forma de lidar com isso é modelar os zeros como um caso particular do nosso modelo GLM.

Volte em sua análise exploratória e em seu ajuste Poisson. Compare a quantidade de zeros presente nos dados e a quantidade de zeros que o modelo de Poisson estima. O modelo de Poisson modela de forma satisfatória os zeros?

Pesquise sobre o modelo de Poisson inflado de zeros, descreva suas possíveis parametrizações e ajuste este modelo aos dados[2]. Com os 3 modelos ajustados (`Poisson`, `NegBin` e `ZeroInflPoisson`) compare métricas de bondade do ajuste –justifique as métricas escolhidas – dos 3 modelos assim como os parâmetros estimados.

Qual modelo você escolheria para fundamentar suas análises?

Referências

[Cameron and Trivedi, 1990] Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of econometrics*, 46(3):347–364.

[2] **Dica**: nosso repositório tem links relevantes.