Modelos para dados de contagem

O modelo de Poisson

Sumário

Dados de Contagem

- Podemos estar interessados em modelar dados de contagem.

- Exemplos
  - Número de chamadas telefônicas por dia em uma call center;
  - Número de acidentes em uma estrada por dia;
  - Número de surtos epiléticos por paciente em dois anos;
  - Número de partos cesariais por hospital/ano;
  - Número de clientes chegando ao caixa de um supermercado;
  - Número de gols por time na primeira rodada do campeonato brasileiro.

- Por que não podemos usar o modelo de regressão estudado anteriormente?
  - $\rightarrow$ Suposição de Normalidade!

- Dados de contagem podem ser modelados por uma distribuição de Poisson.
  - **Discreta** e representa a probabilidade de que um evento ocorra um número especificado de vezes em um intervalo de tempo (espaço).

A distribuição de Poisson

- Seja $Y \sim \text{Poisson}(\lambda)$, então

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad y = 0, 1, \cdots$$

$$\boldsymbol{E(Y) = Var(Y) = \lambda}$$

- A taxa média de ocorrência (λ) é constante ao longo do tempo.

- A informação sobre o número de ocorrências em um período nada revela sobre o número de ocorrências em outro período.

O modelo de regressão de Poisson

- Temos que a variável resposta Y representa dados de contagem ou taxas e $\boldsymbol{X}$ é o vetor de covariáveis.
- **Objetivo:** explicar a variação de Y através de $\boldsymbol{X}$.

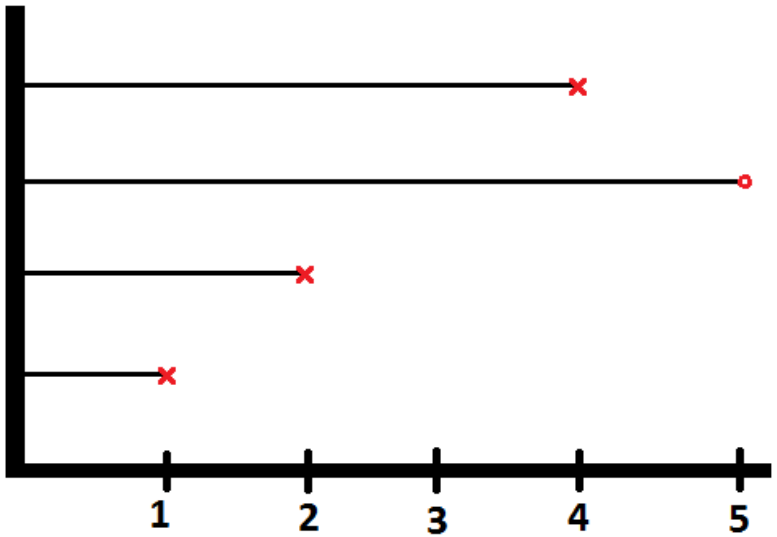
Taxa de incidência

$$TI = \frac{\text{Número de } \textbf{casos novos} \text{ em determinado período}}{\text{Quantidade de pessoa-tempo}}$$

Quantidade de pessoa tempo:tempo em que a população esteve sob risco de desenvolver a doença, sendo o tempo da população igual a soma dos tempos de observação de cada indivíduo.

Exemplos de Taxa de Incidência

Cidade X no período de 5 anos



Vamos calcular a taxa de incidência:

$$TI = \frac{3}{1+2+4+5}$$
$$= \frac{3}{12}$$
$$= 0,25 \text{ por ano}$$

Exemplos de Taxa de Incidência

Mortalidade por gênero de paciente com a doença Y

	Homens	Mulheres	Total
Casos	90	131	221
Pessoas-ano	2465	3946	6911

$$TI_H = \frac{90}{2465} = 0,0365/\text{ano}$$

$$TI_M = \frac{131}{3946} = 0,0332/\text{ano}$$

$$RII = \frac{0,0365}{0,0332} = 1,099$$

Taxa de Incidência

Por que Taxa de Incidência é importante?

Unidades amostrais podem ser acompanhadas por diferentes períodos de tempos.

- Seja y_{ij} a contagem do número de câncer de pele para a i-ésima faixa etária na cidade j.

- Pergunta: A taxa de câncer de pele, ajustada por idade, difere nas diferentes cidades?

- A **regressão de Poisson** seria o modelo **adequado** para modelar a taxa de incidência de eventos (contagens).

Inclusão de covariáveis

- **Como incluir covariáveis?**
- Vamos supor uma amostra de tamanho n.

$$E(Y_i) = \lambda_i(x) \quad i = 1, \cdots, n \text{ e } \lambda_i \geq 0$$

- Vamos usar uma função de ligação logarítimica:

$$\log\left(E(Y_i)\right) = \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Ou seja:

$$\lambda_i = e^{\beta_0} \times e^{\beta_1 x_{i1}} \ldots e^{\beta_p x_{ip}}$$

- Vantagem: garantimos que $\hat{\lambda}_i \geq 0$.

- Como modelamos a taxa?

$$\text{Taxa } = \frac{\lambda(\boldsymbol{x})}{c}$$

onde c é a exposição (tempo, número, área, volume, etc)

- Com a função de ligação logarítimica:

$$\log\left(\frac{\lambda(\boldsymbol{x})}{c}\right) = \boldsymbol{x}'\beta$$
$$\rightarrow \log(\lambda(\boldsymbol{x}) = \log(c) + \boldsymbol{x}'\beta$$

- No R temos o comando *offset* para lidar com o $\log(c)$ (constante sem coeficiente de regressão).

Interpretação dos parâmetros

- Note que agora estamos considerando:

$$\log(\text{contagem ou taxa}) = \boldsymbol{x}'\boldsymbol{\beta}$$

- Os parâmetros não possuem a mesma interpretação do modelo de regressão Normal.

- Fixando $x_2, \cdots, x_p$, quando passamos x_1 de 0 para 1 temos:

$$x_1 = 0 \rightarrow \log(\text{taxa}) = \beta_2 x_2 + \cdots + \beta_p x_p$$
$$x_1 = 1 \rightarrow \log(\text{taxa}) = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Interpretação dos parâmetros

- Então:

$$\log RT = \log\left(\frac{taxa_1}{taxa_0}\right) = \beta_1$$

- Vamos supor que $\exp(\beta_1) = 2$. No caso em que modelos a taxa de incidência temos:

$$RT = \exp(\beta_1) = 2$$

- Isso significa que a taxa de incidência 1 é duas vezes a taxa de incidência de 0.
- E no caso em que modelamos a contagem?
- A interpretação é similiar: a incidência de câncer de 1 é duas vezes a de 0, por exemplo.

Exemplo 1

Partos cesarianos por ano em 20 hospitais (4 privados e 16 públicos)

Partos	Hospitais	cesáreas	Partos	Hospitais	cesáreas
236	0	8	357	1	10
739	1	16	1080	1	16
970	1	15	1027	1	22
2371	1	23	28	0	2
309	1	5	2507	1	22
679	1	13	138	0	2
26	0	4	502	1	18
1272	1	19	1501	1	21
3246	1	33	2750	1	24
1904	1	19	192	1	9

Exemplo 1

Partos cesarianos por ano em 20 hospitais (4 privados e 16 públicos)

- Observe que podemos modelar tanto o número de cesáreas (contagem) quanto a proporção (taxa).
- Seja Y_i o número de cesáreas.
- Suponha que Y_i *Poisson*(μ_i).
- Vamos ajustar $\log(\mu_i) = \beta_0 + \beta_1 \times \text{Partos} + \beta_2 \times \text{Hospital}$.

Exemplo 1

Partos cesarianos por ano em 20 hospitais (4 privados e 16 públicos)

- Assim para $\log(\mu_i) = \beta_0 + \beta_1 \times \text{Partos} + \beta_2 \times \text{Hospital}$, temos:

	estimativa
intercepto	1,351
partos	0,0003261
hospital(1)	1,045

$$\log(\mu) = 1,351 + 0,0003261 \times \text{Partos} + 1,045 \times \text{Hospital}$$

$$E(Y) = e^{1,351} \times e^{0,0003261 \times \text{Partos}} \times e^{1,045\text{Hospital}}$$

Exemplo 1

Partos cesarianos por ano em 20 hospitais (4 privados e 16 públicos)

- Interpretando o parâmetro β_2:

$$e^{1,045} = 2,84$$

- O número de partos por cesárea em hospitais públicos é 2,84 vezes o número de partos por cesárea em hospitais particular.

Exemplo 1

Partos cesarianos por ano em 20 hospitais (4 privados e 16 públicos)

- Note que:

  Público: $\mu_i = \exp(1.351) \times \exp(0,0003261 \times 1000) \times \exp(1,045) = 15,2$

  Particular: $\mu_i = \exp(1.351) \times \exp(0,0003261 \times 1000) = 5,35$

- Isso implica uma média de 15,2 cesáreas a cada 1000 partos em hospitais públicos e 5,4 por 1000 partos em particulares.

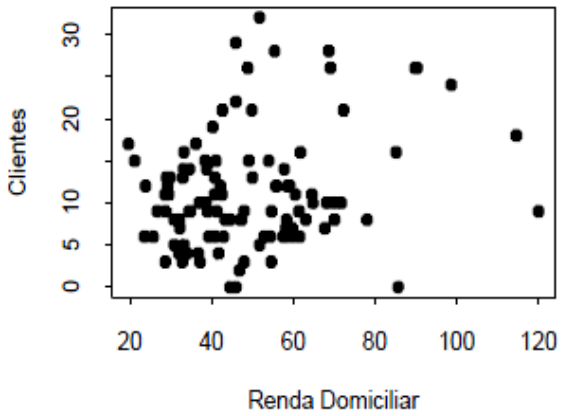
Exemplo 2
Perfil de Clientes (Retirado dos slides do Gilberto Paula)

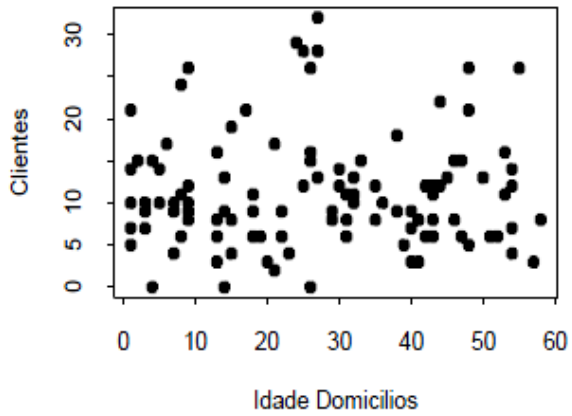
- Dados sobre o perfil dos clientes de uma determinada loja, que foram divididos em 110 áreas de uma cidade.

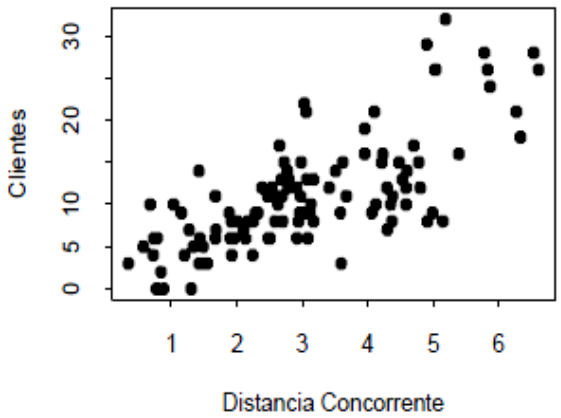
- O número de clientes de cada área que foram à loja num período fixo serão relacionados com as seguintes variáveis em cada área (Neter et al., 1996, p. 613):
  - número de domicílios (em mil) (x_1);
  - renda média anual domiciliar (em mil US$) ($x_2$);
  - idade média (em anos) dos domicílios (x_3);
  - distância ao concorrente mais próximo (x_4);
  - distância à loja (em milhas) (x_5).

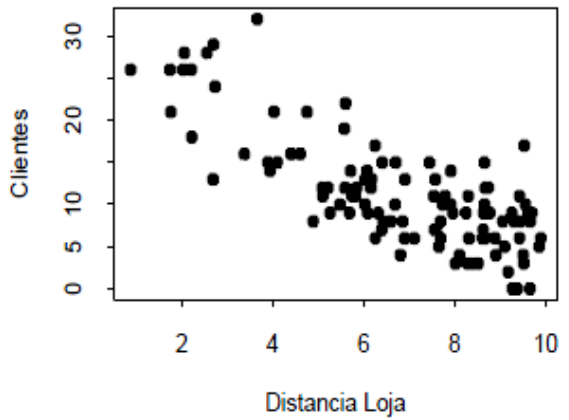
Exemplo 2

Perfil de Clientes (Retirado dos slides do Gilberto Paula)









- Y_i: número de clientes da i-ésima área que foram à loja no período determinado.

- Suponha que $Y_i \sim P(\lambda_i)$, onde:

$$\log(\lambda_i) = \beta_0 + \beta_1 \times x_1 + \cdots + \beta_5 \times x_5$$

Estimando os parâmetros do modelo, encontramos:

Efeito	Estimativa	E.Padrão	z-valor
Constante	2,942	0,207	14,21
Domicilio	0,606	0,142	4,27
Renda	-0,012	0,002	-5,54
Idade	-0,004	0,002	-2,09
Dist1	0,168	0,026	6,54
Dist2	-0,129	0,016	-7,95

- Olhando apenas a tabela, podemos perceber que:
  - O número esperado de clientes na loja cresce com o aumento do número de domicílios na área;
  - O número esperado de clientes na loja diminui com o aumento da renda média e da idade média dos domicílios bem como da distância da área à loja;

- Por exemplo, se aumentarmos em um ano a idade média dos domicílios:

$$\exp(-0,004) = 0.996$$

Assim, esperamos que o número de clientes que irão à loja irá diminuir em 0.4%.

Exemplo 3
Câncer de pele em duas cidades em 1994

Tabela: Dados

	Minneapolis		Dallas	
Idade	Casos	Pop.	Casos	Pop.
15-24	1	172675	4	181343
25-34	16	123065	38	146207
35-44	30	96216	119	121374
45-54	71	92051	221	111353
55-64	102	72159	259	83004
65-74	130	54722	310	55932
75-84	133	32185	226	29007
85+	40	8328	65	7538

Exemplo 3
Câncer de pele em duas cidades em 1994

```
m <- glm(casos ~ idade + cidade + offset(log(pop)),
  family=poisson)
```

	Estimativa	E.P.	p-valor
Intercepto	-10,35	0,096	<0,01
Cidade	0,82	0,052	<0,01
Faixa	0,06	0,0013	<0,01
Cidade*faixa			0,044

- Interpretação (taxa de incidência) para cidade: $\exp(0,82) = 2,2705$

Superdispersão

- Sabemos que se $Y \sim Poisson(\lambda)$ então $E(Y) = Var(Y) = \lambda$.
- Superdispersão ocorre quando há uma inedequação do Modelo de Regressão de Poisson.
- Dizemos que houve superdispersão quando $Var(Y) > E(Y)$

1. Função de ligação inadequada. Nós vimos o caso da função de ligação logarítimica.

2. Não inclusão de covariáveis importantes no preditor linear:
   - Desconhecidas;
   - Não foram medidas.

3. Excesso de zeros:
   - Comumente existem situações com excesso de contagens zero;
   - horários inadequados, pessoas não contaminadas, entre outros.

Possível Solução

Incluir mais um parâmetro na modela para incorporar essa "extra variação"