Análise de Dados Categóricos

Modelo de Regressão de Poisson

Enrico A. Colosimo/UFMG

http://www.est.ufmg.br/~enricoc/

Departamento de Estatística
Universidade Federal de Minas Gerais

Revisão: Modelos Lineares Generalizados

Modelos Lineares Generalizados (MLG) é uma classe unificada de modelos de Regressão.

1. Considere $Y_1, \ldots Y_n$ uma amostra aleatória de respostas univariadas.

2. Um vetor de p-covariáveis associados a cada resposta Y_i. Ou seja

$$X_i = \begin{pmatrix} X_{i0} \\ X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}$$

em que $X_{i0} = 1$; $i = 1, \ldots, n$.

3 O MLG é definido por três componentes:

- Distribuição de Y_i.
- Componente Sistemático (preditor linear).

$$\eta_i = X_i' \beta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Função de Ligação.

Modelos Lineares Generalizados (MLG)

1. A distribuição de Y_i pertence a família exponencial .
2. Inferência MLG
   - EMV / Método Escore de Fisher;
   - Estatísticas Assintóticas: Wald, RV, escore;
   - Adequação do modelo
     - Testes: estatísticas do desvio (TRV) e de Pearson e seus, respectivos, resíduos;
     - Gráfico: envelope.

Modelo de Poisson: Resposta Contagem

- Interesse em modelar resposta do tipo contagem (ou taxa).
- Exemplos
  - Número de chamadas telefônicas por dia em um call center;
  - Número de acidentes em uma estrada por mês;
  - 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 por hora;
  - Número de gols por time na primeira rodada do campeonato brasileiro;
  - Número de ovos de um parasita por mm^3 de fezes;
  - etc, etc.

Resposta: Contagem

1. Por que não devemos usar o modelo de regressão linear?
   - $\rightarrow$ Suposição de Normalidade!
   - $\rightarrow$ Suposição de Homocedasticidade!
2. Soluções
   - Usar transformação na resposta (por exemplo, raiz quadrada).
   - Usar mínimos quadrados ponderados.
   - Mais indicado: usar modelo/distribuição de Poisson.

A distribuição de Poisson

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

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

- Pertence a família exponencial.
- O número médio de ocorrência (λ) é constante ao longo do tempo.
- Incrementos independentes: a informação sobre o número de ocorrências em um período nada revela sobre o número em outro período distinto.

Propriedades da distribuição de Poisson

1. Seja $Y \sim$ Binomial (n, p), então, se

$$np \to \lambda \quad \text{e} \quad p \to 0$$

$Y \to$ Poisson (λ)

Exemplos:

- Incidência de uma forma rara de câncer em pequenas regiões geográficas.
- Tratar a resposta Y como binomial ou contagem?

Propriedades da distribuição de Poisson

2 Incrementos independentes e taxa média de ocorrência constante.

Exemplos:

- Número de chamadas telefônicas por dia em um call center;
- Número de acidentes em uma estrada por mês;
- Número de clientes chegando ao caixa de um supermercado por hora;
- É razoável modelar Y como Poisson?

Propriedades da distribuição de Poisson

3. Soma de distribuições de Poisson independentes tem uma distribuição de Poisson com parâmetro que é a soma das taxas individuais.

Esta propriedade pode ser importante em situações que temos somente informação de contagens agregadas.

Propriedades da distribuição de Poisson

4 Contagens muito grandes.

- Regra Empírica: uma aproximação normal é justificável, o que possibilita utilizar o modelo de regressão linear.
- Neste caso, uma transformação raiz quadrada estabiliza a variância.

Propriedades da distribuição de Poisson

5 Distribuição de Poisson surge naturalmente quando o tempo entre eventos é independente e identicamente distribuído com distribuição exponencial.

Este fato é equivalente a incrementos independentes e taxa média constante.

Propriedades da distribuição de Poisson

6. Exposição de indivíduos diferentes em estudos longitudinais.

Exemplos:

- Número de surtos epiléticos por paciente.
- Número de internações por paciente.
- Cada paciente foi acompanhado por um perído diferente de tempo.

Modelar taxa ao invés de contagem (usar offset no modelo).

EMV e Propriedades Assintóticas/Exatas

1. Amostra homogênea de tamanho n.
2. $L(\lambda|dados) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}$
3. EMV:

$$\widehat{\lambda} = \bar{y}$$

e

$$Var(\widehat{\lambda}) = \frac{\lambda}{n}$$

O modelo de regressão de Poisson

- Temos que a variável resposta Y representa uma contagem ou taxa e $\mathbf{X}$ é o vetor de covariáveis.
- **Objetivo:** explicar a variação de Y através de $\mathbf{X}$.
- Tipo de Estudo:
  - Transversal: Y: contagem /unidade.
  - Longitudinal: Y: taxa = contagem/tempo.

Estudo Longitudinal

Indivíduos ou pacientes acompanhados por diferentes períodos.
Exemplos: número de surtos epiléticos por paciente.

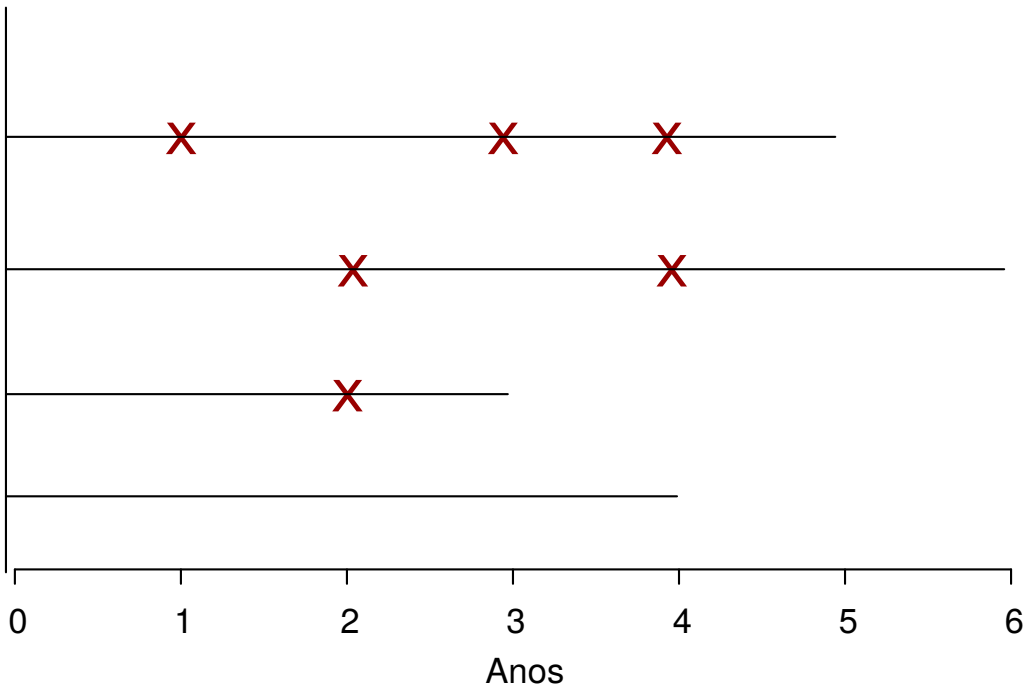
Taxa de incidência

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

- Quantidade de pessoa tempo: tempo em que a população esteve sob risco de desenvolver o evento
- O tempo da população é igual a soma dos tempos de observação de cada indivíduo.

Exemplos de Taxa de Incidência

Paciente no período de 6 anos



Vamos calcular a taxa de incidência:

$$TI = \frac{6}{4+3+5+6}$$
$$= \frac{6}{18}$$
$$= 0,33 \text{ por ano}$$

Exemplos de Taxa de Incidência -
Episódios de diárreia em crianças por semana

	Meninos	Meninas	Total
Ocorrências	90	131	221
Pessoas-semana	2465	3946	6911

$$TI_O = \frac{90}{2465} = 0,0365/\text{semana}$$

$$TI_A = \frac{131}{3946} = 0,0332/\text{semana}$$

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

Taxa de Incidência

Por que Taxa de Incidência é importante?

Unidades amostrais podem ser expostas/acompanhadas por diferentes períodos de tempos.

Exemplo:

1. Seja Y o número de surtos epilépticos por paciente em diferentes cidades.
   - Os pacientes na amostra foram expostos/acompanhados por diferentes períodos de tempo.
   - **Pergunta:** A taxa de surtos epilépticos, ajustada por idade do paciente, difere nas diferentes cidades?

2. Seja Y o número de câncer de pele em uma certa faixa etária na população alvo.
   - Os indivíduos variam por faixa etária na população e por tempo de acompanhamento.
   - **Pergunta:** A taxa de câncer de pele difere nas diferentes classes etárias?

3. O modelo de **regressão de Poisson** é o indicado para modelar a taxa de incidência de eventos (contagens).

Inclusão de covariáveis

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

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

- Vamos usar uma função de ligação logarítmica:

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

Ou seja:

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

- Vantagem: garantimos que $\hat{\lambda}_i \geq 0$ e é a ligação canônica da família exponencial.

- Como modelamos a taxa de incidência?

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

em que c é a medida de exposição (tempo, número, área, volume, etc)

- Com a função de ligação logarítmica:

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

- $\log(c)$ é chamado de *offset*.
- 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 modelamos a taxa de incidência temos que:

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

- Isso significa que a taxa de incidência para x=1 é duas vezes a taxa de incidência para x=0.
- E no caso em que modelamos a contagem?
- A interpretação é similiar: a ocorrência média do evento para x=1 é duas vezes a x=0.

Inferência para β

1. EMV para uma amostra de tamanho n
2. $$L(\lambda|dados) = \prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$
3. $\lambda_i = \exp(X_i\beta)$.
4. Função Escore
5. Matriz de Informação

Adequação do Modelo

H_0 : o modelo é adequado.

1. Estatística Qui-quadrado
2. $$X^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

$$\hat{y}_i = \exp(\text{exposição}_i + \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_p x_{ip})$$

3. E os componentes de X^2, que são os resíduos.
4. Estatística do Desvio
5. $$D = -2\{l(\text{modelo corrente}) - l(\text{modelo saturado})\}$$

$$D = 2\sum_{i=1}^{N} (y_i \log(y_i/\hat{y}_i) - (y_i - \hat{y}_i))$$

Continuação: Adequação do Modelo

H_0 : o modelo é adequado.

1. Estatística Qui-quadrado X^2 e D têm sob H_0 uma distribuição qui-quadrado com $N - p$ graus de liberdade.

2. No entanto a afirmação acima somente é verdade se $N << n$, em que N é o número de diferentes combinações dos valores das covariáveis.

3. Devemos ter cuidado com a distribuição de X^2 e D. Na presença de covariáveis contínuas ou tamanho de amostra pequeno ($N \approx n$), não tem distribuição qui-quadrado.

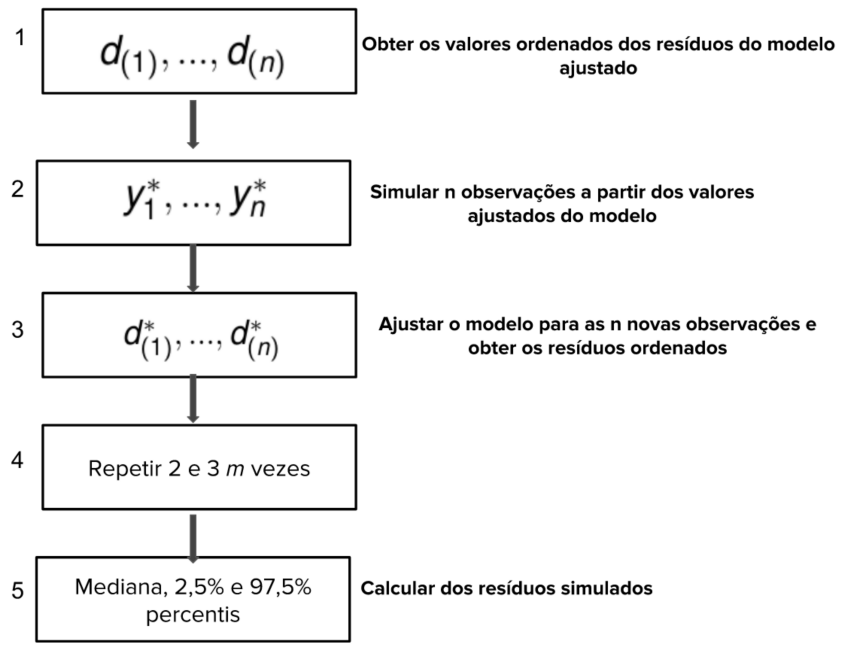
Adequação do Modelo

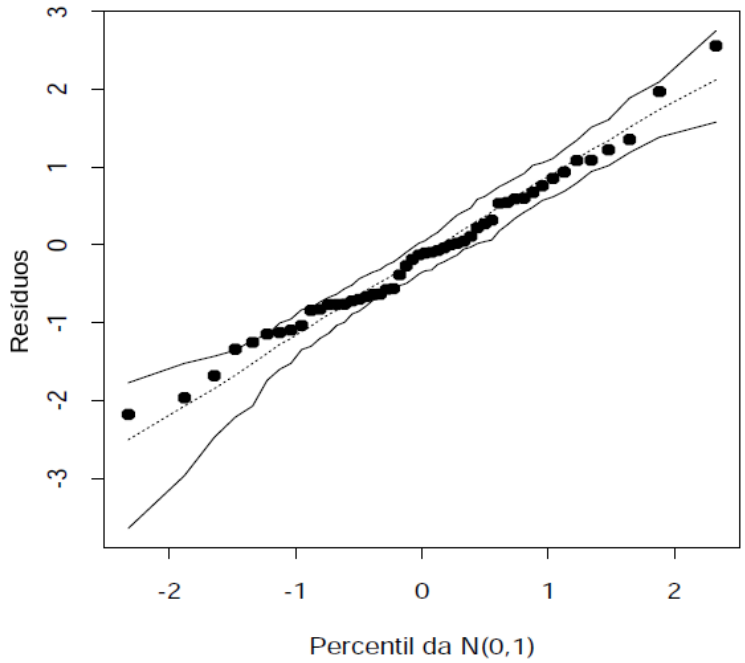
Gráfico de Envelope

- O gráfico de probabilidade normal com envelope simulado é usado para verificar adequação do modelo;
- Atkinson (1981) propôs a construção por simulação de Monte Carlo de uma banda de confiança para os resíduos (modelo normal)
- Williams (1987) discute a construção dos envelopes para os MLG's.
- No caso dos MLG's, a construção é feita com os resíduos gerados do modelo ajustado;

Gráfico de envelope

Passos para a construção





- Resíduos deviance;
- usualmente, $m = 19$ sugerido por Atkison (1981);
- quantis teóricos da distribuição normal x resíduos do modelo inicial;
- a mediana e os percentis dos resíduos simulados formam o envelope

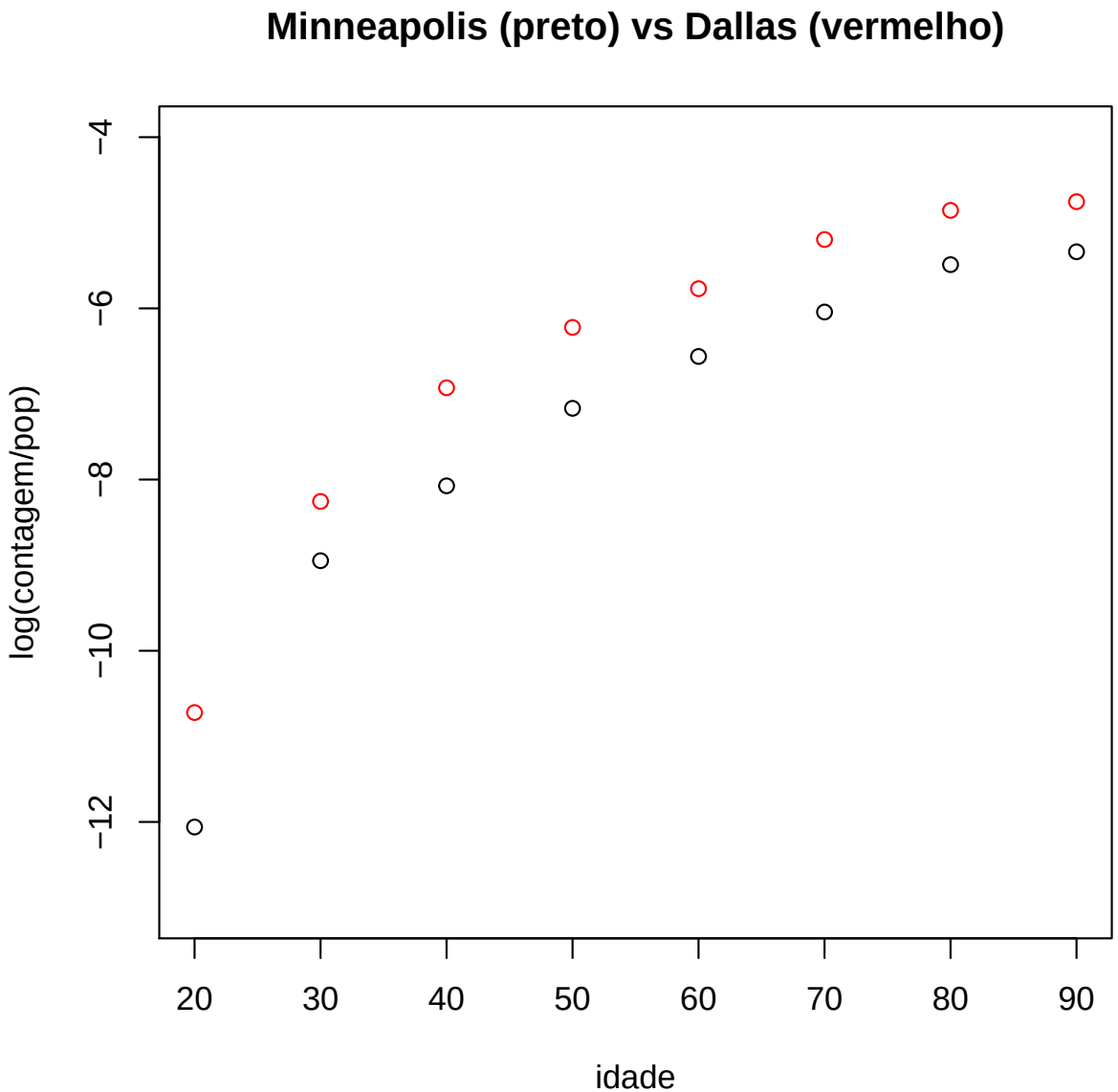
Exemplo 1

Câncer de pele não melanoma 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

Exemplo1: Câncer de pele em duas cidades em 1994



Exemplo 1

Câncer de pele em duas cidades em 1994

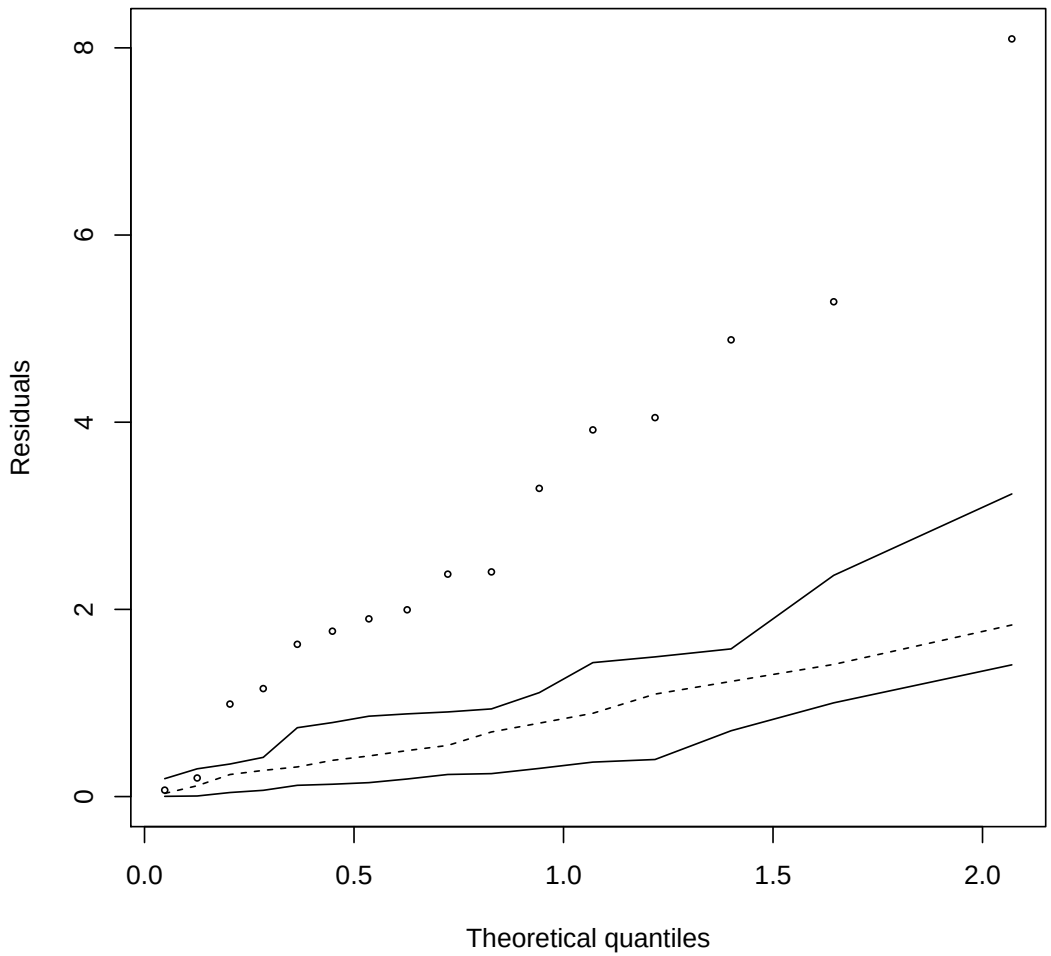
```
m <- glm(casos ~ idade + cidade + offset(log(pop)),
 family=poisson)
```

valor-p (qui-quadrado e desvio) $<$ 0,001.

Este modelo não é adequado pois a idade não tem um comportamento linear na escala de log(taxa).

Exemplo 1

Envelope - Modelo Inadequado



Exemplo 1

Câncer de pele em duas cidades em 1994

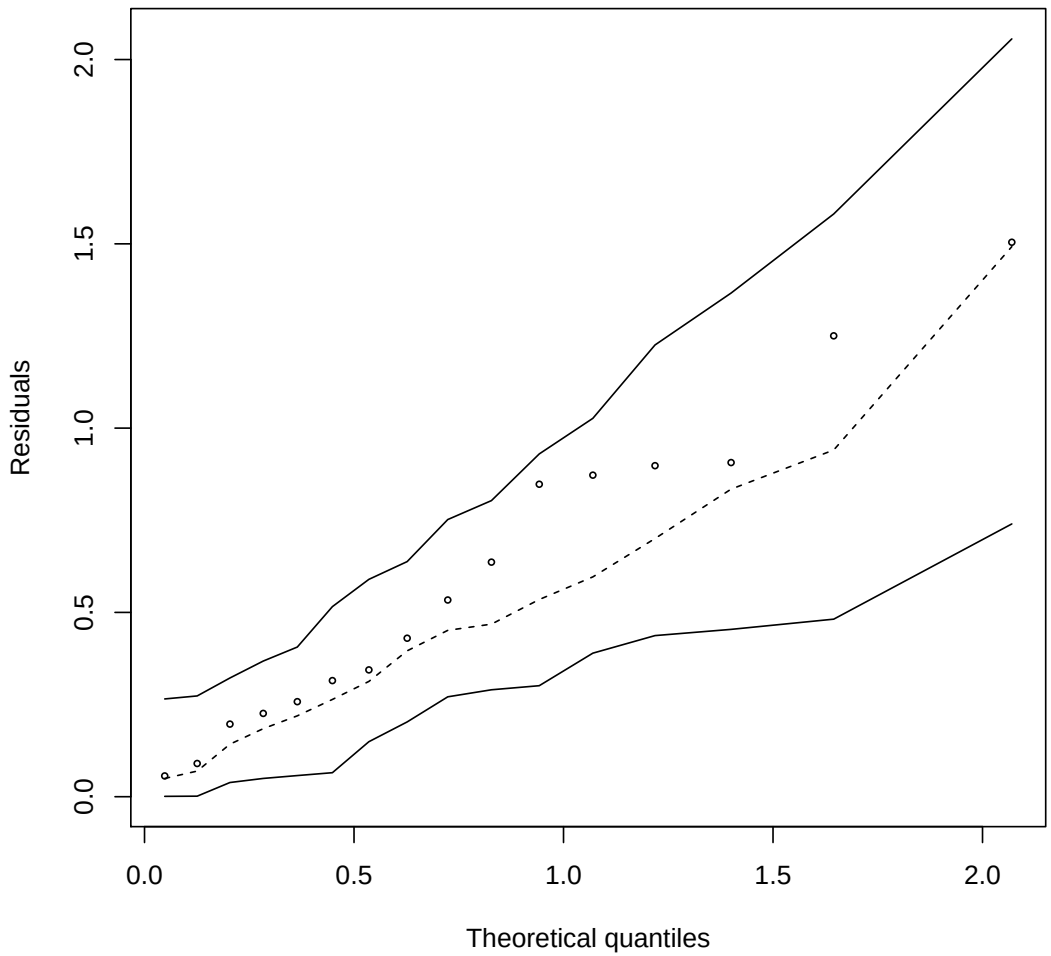
```
m <- glm(casos ~ factor(idade) + cidade + offset(log(pop))
 family=poisson)
```

Valor-p (deviance) = 0,316 e Valor-p (qui-quadrado) = 0,707;

- Interpretação (taxa de incidência) para cidade: $\exp(0,804) = 2,23$ (IC 95%; 2,0; 2,5). Ou seja, a taxa de incidência de câncer de pele em Dallas é 2,2 vezes a de Minneapolis.
- A taxa de incidência de câncer de pele aumenta com o aumento da idade.

Exemplo 1

Envelope - Modelo Adequado



Superdispersão

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

Superdispersão
Possíveis Causas

1. Função de ligação inadequada.

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.
   - Lambert (1992).

Superdispersão
Solução

Possível Solução

Incluir mais um parâmetro no modelo para incorporar essa "extra variação"

Usar o modelo binomial negativo (mais utilizado).

Modelo de Regressão Binomial Negativo

Vamos supor uma amostra de tamanho n.

$$E(Y_i|x_i, \tau_i) = \mu_i \tau_i$$

em que, τ_i representa a heterogeneidade não observada.

$$E(Y_i|x_i, \tau_i) = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \exp(\epsilon_i)$$

Ou seja:

$$\tau_i = e^{\epsilon_i}$$

Modelo de Regressão Binomial Negativo

$$p(y_i|x_i, \tau_i) \sim Poisson(\mu_i \tau_i)$$

e

$$f(\tau_i) \sim gama(\alpha, \alpha)$$

Então

$$p(y_i|x_i) \sim \text{Binomial Negativa}$$

Isto significa que,

$$Var(Y_i|x_i) = E(Y_i|x_i)(1 + \delta)$$

em que $\delta = \alpha\mu_i > 0$

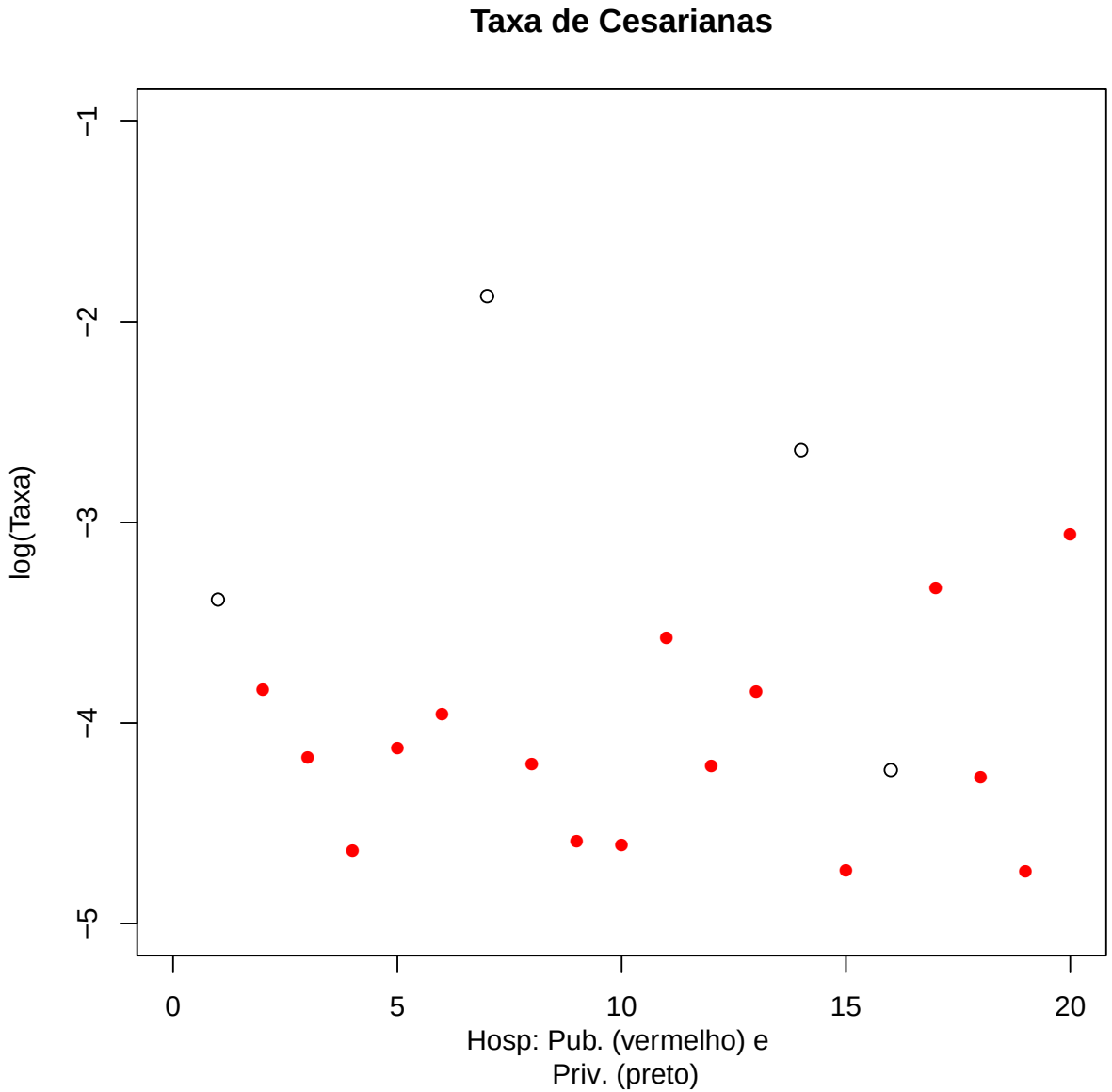
Exemplo 2

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

Partos	Hospitais	cesáreas
236	0	8
739	1	16
970	1	15
2371	1	23
309	1	5
679	1	13
26	0	4
1272	1	19
3246	1	33
1904	1	19

Partos	Hospitais	cesáreas
357	1	10
1080	1	16
1027	1	22
28	0	2
2507	1	22
138	0	2
502	1	18
1501	1	21
2750	1	24
192	1	9

Exemplo 2- Partos cesarianos por ano em 20 hospitais



Exemplo 2

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

- Devemos modelar a proporção de cesáreas (taxa).
- Seja Y_i o número de cesáreas.
- Suponha que $Y_i \sim Poisson(\lambda_i)$.
- Vamos ajustar $\log(\lambda) = \log(\text{Partos}) + \beta_0 + \beta_1 \times \text{Hospital}$.

Exemplo 2

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

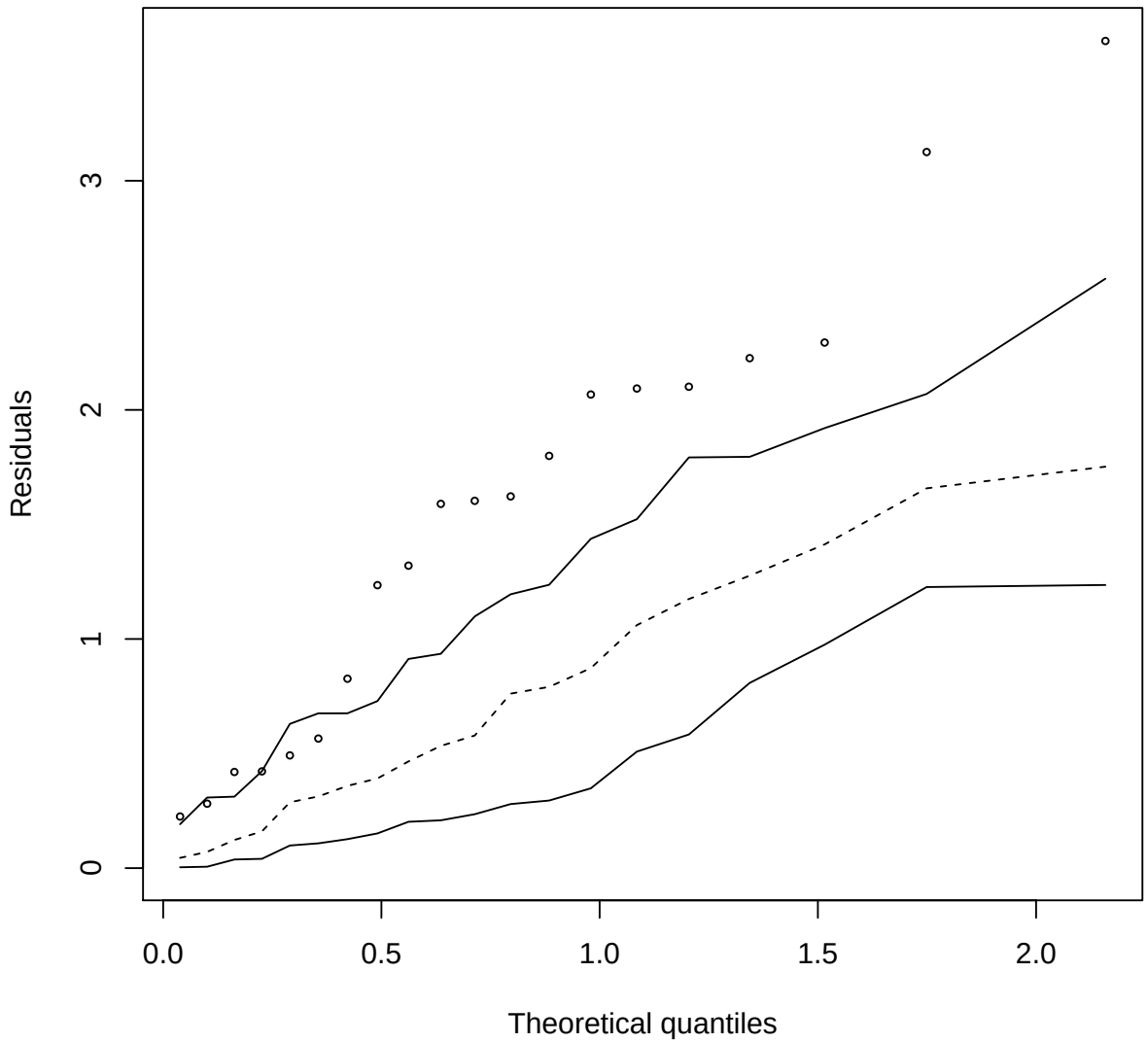
- Assim para $\log(\lambda_i) = log(\text{Partos}_i) + \beta_0 + \beta_1 \times \text{Hospital}_i$, temos:

	estimativa
intercepto	-3,29
hospital(1)	-1,03

$$\log(\widehat{\lambda}_i) = log(\text{Partos}_i) + 3,29 - 1,03 \times \text{Hospital}$$

Modelo não é adequado (valor-p $< 0,001$ para D e X^2).

Exemplo 2: Partos cesarianos por ano em 20 hospitais



Exemplo 2

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

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

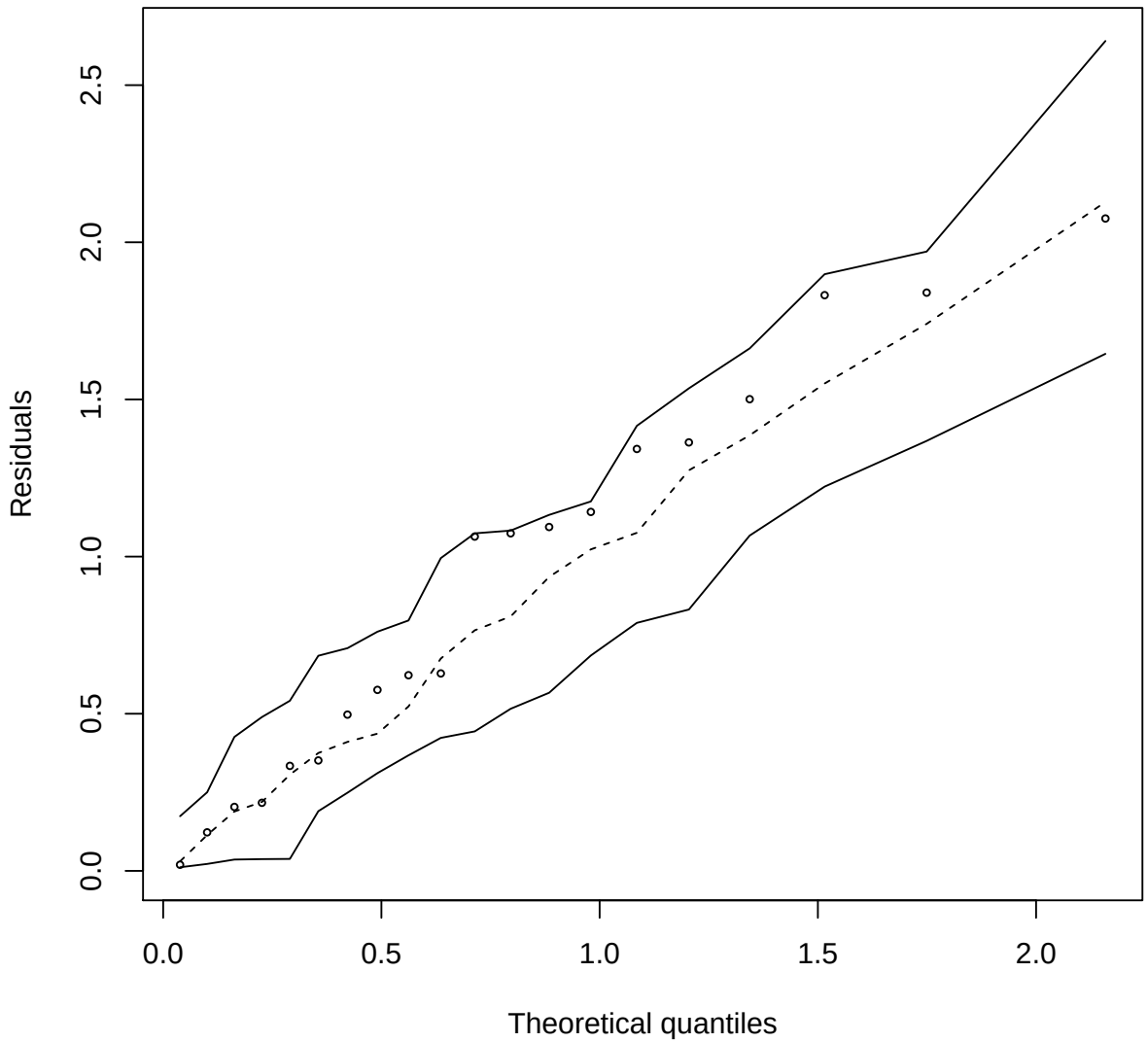
Modelo Binomial Negativo

	estimativa
intercepto	-3,12
hospital(1)	-0,99

Modelo é adequado (valor-p = 0,174 (desvio) e valor-p=0,0482 (qui-quadrado)).

Interpretação: $1/exp(-0,988) = 2,7$ (IC 95%; 1,4; 5,4), a ocorrência de cesarianas em hospitais privados é 2,7 vezes a de públicos.

Exemplo 2: Partos cesarianos por ano em 20 hospitais



Uum Breve Roteiro para a Análise de Dados

1. Entender o Estudo.
2. Descrever o Estudo: Importância e Objetivos.
3. Identificar o desenho amostral.
4. Exploração e Verificação da Consistência do Banco de Dados (cada variável separadamente).
5. Análise bivariada: resposta com cada uma variável separadamente.

REGRA EMPíRICA (na presença de várias covariáveis):

Excluir covariáveis com valor-p $> 0,25$ no passo anterior.

Uum Breve Roteiro para a Análise de Dados

6 Modelo de Regressão (Poisson/Logística)

- Utilizar de preferência o Teste da Razão de Verossimilhança;
- Investigar possíveis associações entre as covariáveis (colinearidade);
- Investigar a forma de inclusão de covariáveis contínuas;
- Obter um "Modelo Final"utilizando algum método de construção de modelos.

7 Verificar a adequação do modelo ajustado.

8 Incluir possíveis termos de interação.

9 Interpretar o modelo final apresentando intervalos de confiança para as quantidades de interesse.

10 Escrever o Relatório.

Exemplo 3: Miller Lumber Company: número de clientes (Kutner et. al., 2004)

- Levantamento feito durante duas semanas sobre clientes que visitaram uma certa loja;
- Foi identificado a qual setor censitário cada cliente residia e assim contado o número de clientes em cada setor;
- Todos os setores censitários têm aproximadamente a mesma população;
- Há informações de 110 setores censitários;
- O objetivo do estudo: é verificar a possível associação entre o número de clientes com: número de casas no setor censitário, renda média, idade média das casas , distância até concorrente mais próximo e distância até a loja.

Exemplo 3

Miller Lumber Company Example (Kutner et. al., 2004)

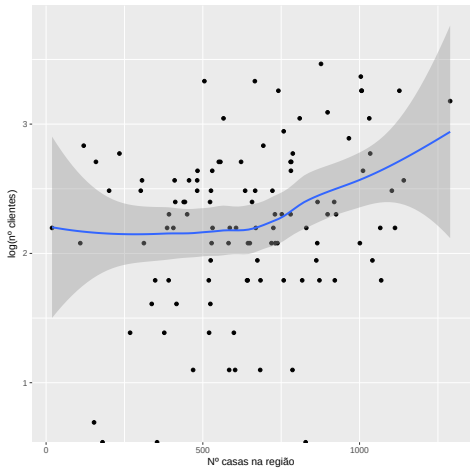
Tabela: Miller Lumber Company Example (Kutner et. al., 2004))

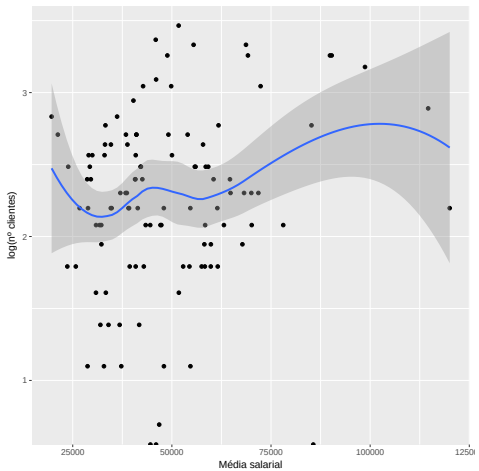
	Housing units	Average income	Average age	Competitor distance	Store distance	Numbers of customers
1	606	41.393	3	3.04	6.32	9
2	641	23.635	18	1.95	8.89	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
109	268	34.022	54	1.20	9.51	4
110	519	52.850	43	2.92	8.62	6

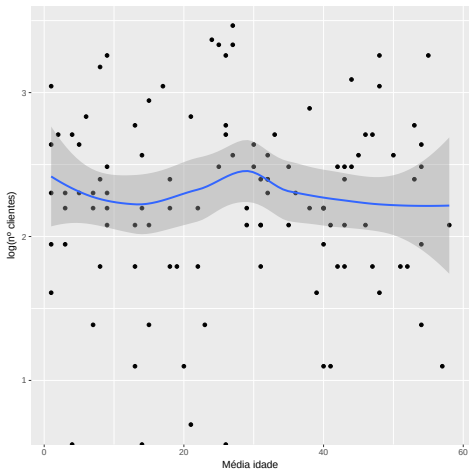
Considere Y_i o número de clientes de cada setor censitário que visitaram a loja e suponha $Y_i \sim Poisson(\lambda_i),\ i = 1, ..., 110$.

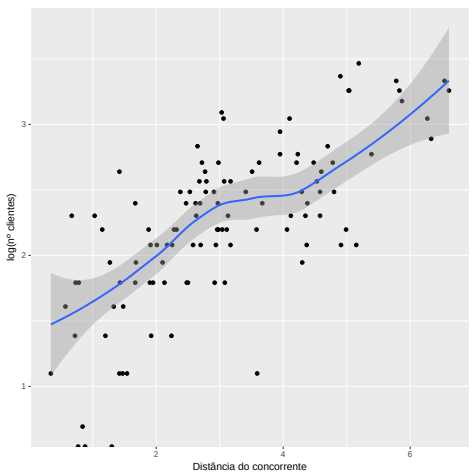
Exemplo 3

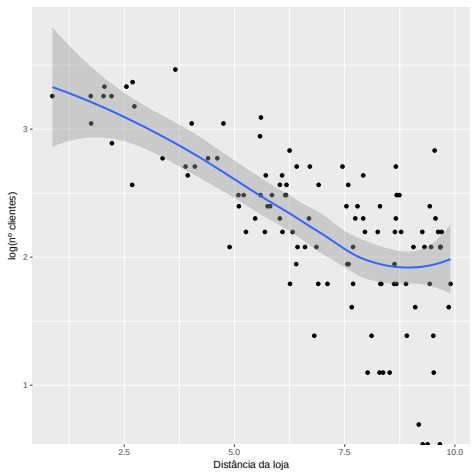
Miller Lumber Company Example (Kutner et. al., 2004)











Exemplo 3

Miller Lumber Company Example (Kutner et. al., 2004)

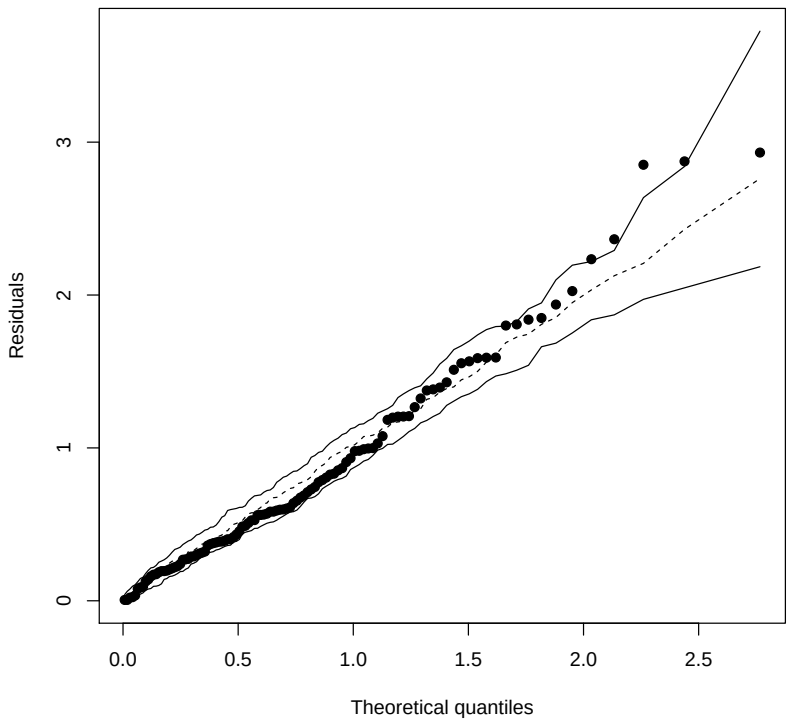
Ajustando $log(\lambda) = log(n^o clientes) = \beta_0 + \beta_1 x2 + \beta_2 x3 + \beta_3 x4 + \beta_4 x5 + \beta_5 x6$ temos o seguinte resultado

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.94244	0.20725	14.19769	< 0.001
No. unidades	0.00061	0.00014	4.26232	< 0.001
Renda média	-0.00001	0.00000	-5.53405	< 0.001
Idade média (casa)	-0.00373	0.00178	-2.09127	0.03650
Distância competidor	0.16838	0.02577	6.53432	< 0.001
Distância à loja	-0.12877	0.01620	-7.94815	< 0.001

Como $D = 114.99$ está próximo aos gl de χ^2 que é $110 - 6 = 104$, indica um ajuste adequado.

Exemplo 3

Miller Lumber Company Example (Kutner et. al., 2004)



Exemplo 3

Miller Lumber Company Example (Kutner et. al., 2004)

- não foi identificado nenhum termo de interação significativo;
- o valor negativo das estimativas de renda salarial média, idade média da casa e distância do setor até a loja, indica que quando aumenta o valor de uma dessas covariáveis, o número médio de clientes diminui;
- o número de clientes que visitam a loja aumenta quando aumenta a distância do concorrente e o número de unidades no setor.
- $exp(10 * 0.00061) = 1.006119$, aumentando 10 casas no setor censitário, aumenta o número de clientes em 0,6%.
- $exp(-0.12877) = 0.8791762$, para cada km que aumentamos na distância à loja, reduzimos em 12% o número de clientes.

Pacote `hnp`

Half-Normal Plots with Simulation Envelopes

- Útil para construir gráficos de resíduos;
- resíduos: deviance (`glm`), student (`aov`, `lm`), pearson (`zeroinfl`, `hurdle`);
- número de simulações MC `default = 99`