

# Exercício 4

## Descrição

No arquivo **dboston.dat** é apresentado parte do conjunto de dados de uma amostra aleatória de 506 setores censitários de 96 distritos da cidade de Boston (USA) em 1970. O objetivo principal do estudo é tentar explicar a relação entre o preço mediano das residências ocupadas pelos proprietários em cada setor censitário com algumas variáveis explicativas. Vamos considerar apenas quatro variáveis explicativas que estão colocadas na seguinte ordem:

- dist: distância ponderada do distrito a cinco centros de emprego de Boston (em milhas)
- tax: imposto distrital anual do imóvel (por 10 mil USD)
- ptratio: relação aluno-professor no distrito
- lstat: porcentagem da população com baixa renda
- medv: preço mediano das residências ocupadas pelos proprietários (em mil USD)

## Enunciado

- Faça inicialmente uma análise descritiva construindo por exemplo boxplots e diagramas de dispersão de cada variável explicativa contra a variável resposta.
- Apresente também a densidade da variável resposta.
- Para cada ligação (logarítmica, identidade e recíproca) proponha um modelo com resposta gama e selecione as variáveis explicativas usando o método AIC.
- Através de procedimentos de diagnóstico escolha um modelo.
- interprete os parâmetros do modelo escolhido.

## Leitura dos dados

```
dboston <- data.frame(scan("dados/dboston.dat",  
                           list(dist=0, tax=0, ptratio=0, lstat=0, medv=0)))
```

	dist	tax	ptratio	lstat	medv
1	4.09	296.00	15.30	4.98	24.00
2	4.97	242.00	17.80	9.14	21.60
3	4.97	242.00	17.80	4.03	34.70
4	6.06	222.00	18.70	2.94	33.40
5	6.06	222.00	18.70	5.33	36.20

## Análise descritiva

Abaixo estão os boxplots (figura 1) de todas as variáveis do banco de dados. Vemos pelos gráficos que as variáveis parecem apresentar distribuição ligeiramente assimétrica.

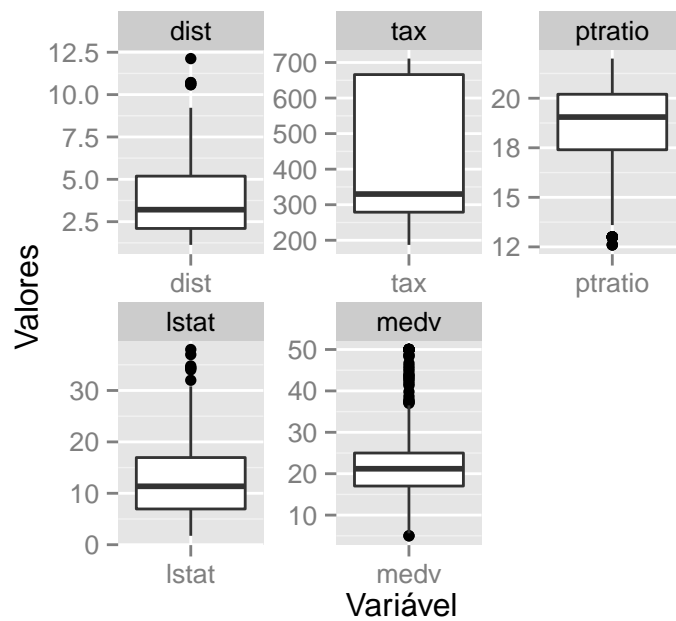


Figure 1: Boxplots das variáveis do banco de dados

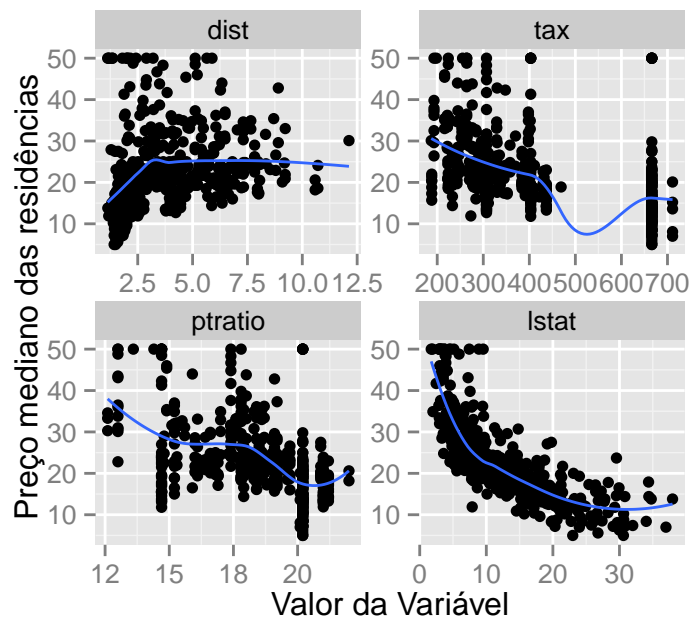


Figure 2: Gráficos de dispersão das variáveis explicativas pela variável resposta

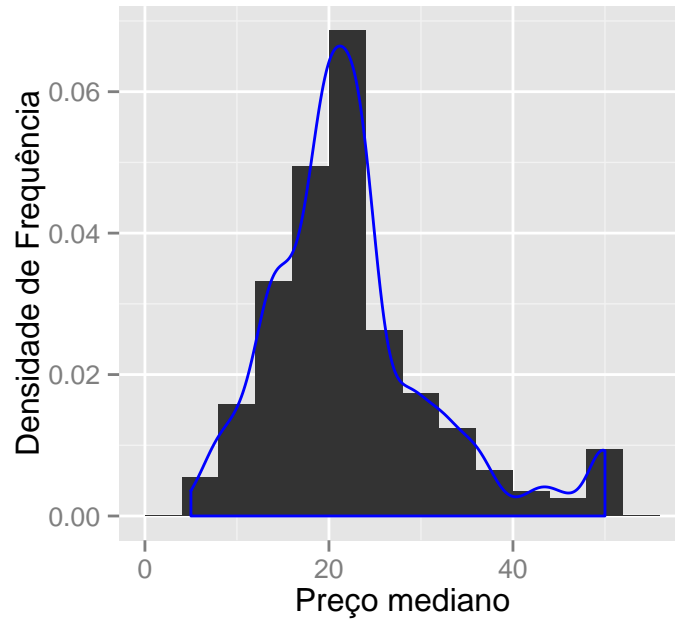


Figure 3: Gráficos de dispersão das variáveis explicativas pela variável resposta

Nos gráficos de dispersão (figura 2) das variáveis pela variável resposta. Principalmente as variáveis **dist** e **lstat** têm relação forte e não linear com o preço dos imóveis. Já as variáveis **tax** e **ptratio** parecem apresentar uma relação linear com o preço mediano.

A densidade da variável resposta **medv** (figura 3) é assimétrica a esquerda, assemelhando-se à distribuição Gama.

## Modelo

### Ligação logarítmica

#### Descrição do modelo

O modelo que será ajustado é de regressão Gamma com ligação logarítmica. Ele é da forma:

$$y_i \sim \text{Gama}(\mu_i, \phi)$$

De forma que:

$$\log(\mu_i) = x_i \beta$$

#### Seleção das variáveis

Como o número de variáveis explicativas é pequeno, vamos ajustar todos os modelos possíveis e escolher aquele que tiver o maior AIC. Os modelos estão apresentados na tabela abaixo.

```
todos.modelos <- fitall(y = dboston$medv, x = dboston[,1:4], method = "glm",
  family = Gamma(link = "log"))
```

variaveis	logL	AIC	BIC	n
(Intercept), dist	-1780	-3566	-3579	506
(Intercept), dist, tax	-1724	-3455	-3472	506
(Intercept), tax	-1724	-3453	-3466	506
(Intercept), ptratio	-1722	-3451	-3463	506
(Intercept), dist, ptratio	-1713	-3434	-3451	506
(Intercept), dist, tax, ptratio	-1690	-3390	-3411	506
(Intercept), tax, ptratio	-1690	-3388	-3405	506
(Intercept), lstat	-1546	-3099	-3112	506
(Intercept), dist, lstat	-1542	-3092	-3109	506
(Intercept), tax, lstat	-1533	-3073	-3090	506
(Intercept), dist, tax, lstat	-1517	-3044	-3065	506
(Intercept), ptratio, lstat	-1509	-3026	-3042	506
(Intercept), tax, ptratio, lstat	-1506	-3022	-3043	506
(Intercept), dist, ptratio, lstat	-1501	-3013	-3034	506
(Intercept), dist, tax, ptratio, lstat	-1492	-2996	-3021	506

O modelo com maior AIC é o modelo com todas as variáveis explicativas do banco de dados: **dist**, **tax**, **ptratio** e **lstat**. Mas pelo princípio de parcimônia vamos usar um modelo com menos variáveis que também tem o AIC alto, este modelo tem as seguintes variáveis: **dist**, **ptratio** e **lstat**.

Este modelo pode ser ajustado no R usando o seguinte comando:

```
modelo <- glm(medv ~ dist + ptratio + lstat, data = dboston,
              family = Gamma(link = "log"))
```

## Análise de Diagnóstico

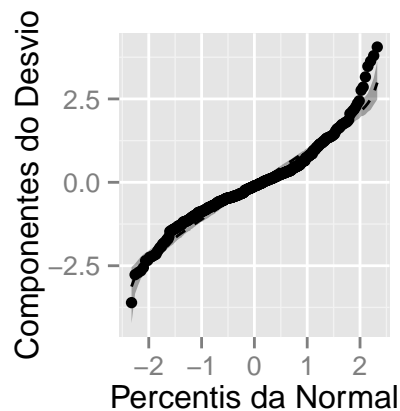


Figure 4: Gráfico Quantil-Quantil

No gráfico quantil-quantil (figura 4) podemos ver que alguns pontos da cauda saem bastante da banda de confiança, indicando que a suposição de distribuição Gamma no modelo não é válida.

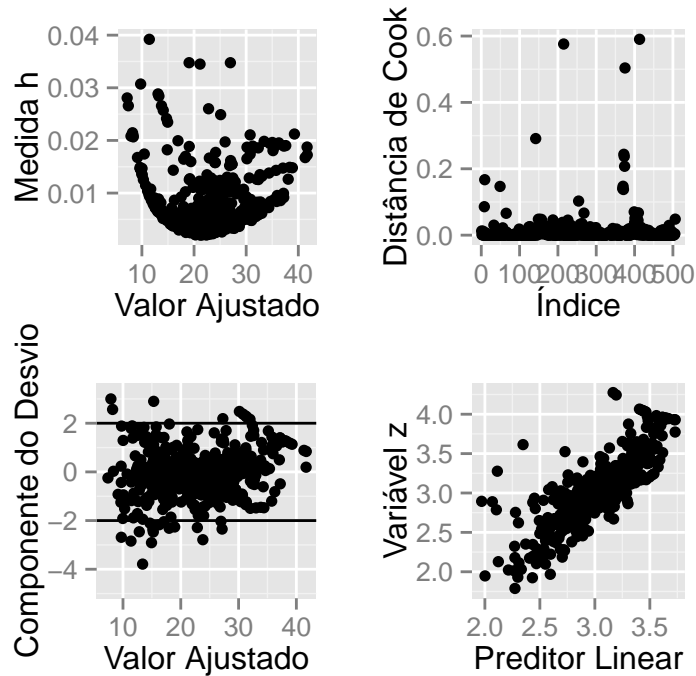


Figure 5: Gráficos de diagnóstico

Nos gráficos da figura 5 vemos não existe nenhum ponto com influência elevada. No gráfico do Valor ajustado pelo componente do desvio, vemos que são poucos os pontos que saem do intervalo  $[-2, 2]$  indicando que o ajuste está razoável. Além disso no gráfico de da variável  $z$  pelo preditor linear verificamos uma tendência linear indicando que a função de ligação escolhida é adequada.

## Ligação identidade

O modelo que será ajustado é de regressão Gamma com ligação logarítmica. Ele é da forma:

$$y_i \sim \text{Gama}(\mu_i, \phi)$$

De forma que:

$$\mu_i = x_i \beta$$

## Seleção das variáveis

Como o número de variáveis explicativas é pequeno, vamos ajustar todos os modelos possíveis e escolher aquele que tiver o maior AIC. Os modelos estão apresentados na tabela abaixo.

```
todos.modelos <- fitall(y = dboston$medv, x = dboston[,1:4], method = "glm",
                        family = Gamma(link = "identity"))
```

variaveis	logL	AIC	BIC	n
(Intercept), dist	-1778	-3563	-3575	506

variaveis	logL	AIC	BIC	n
(Intercept), dist, tax	-1725	-3457	-3474	506
(Intercept), tax	-1725	-3455	-3468	506
(Intercept), ptratio	-1719	-3444	-3457	506
(Intercept), dist, ptratio	-1710	-3427	-3444	506
(Intercept), dist, tax, ptratio	-1692	-3394	-3416	506
(Intercept), tax, ptratio	-1693	-3393	-3410	506
(Intercept), dist, lstat	-1606	-3220	-3237	506
(Intercept), lstat	-1606	-3218	-3231	506
(Intercept), tax, lstat	-1582	-3172	-3189	506
(Intercept), dist, tax, lstat	-1575	-3160	-3181	506
(Intercept), ptratio, lstat	-1562	-3132	-3149	506
(Intercept), dist, ptratio, lstat	-1560	-3131	-3152	506
(Intercept), tax, ptratio, lstat	-1555	-3120	-3142	506
(Intercept), dist, tax, ptratio, lstat	-1549	-3109	-3135	506

O modelo com maior AIC é o modelo com todas as variáveis explicativas do banco de dados: **dist**, **tax**, **ptratio** e **lstat**. Mas pelo princípio de parcimônia vamos usar um modelo com menos variáveis que também tem o AIC alto, este modelo tem as seguintes variáveis: **tax**, **ptratio** e **lstat**.

Este modelo pode ser ajustado no R usando o seguinte comando:

```
modelo <- glm(medv ~ tax + ptratio + lstat, data = dboston,
              family = Gamma(link = "identity"))
```

## Análise de Diagnóstico

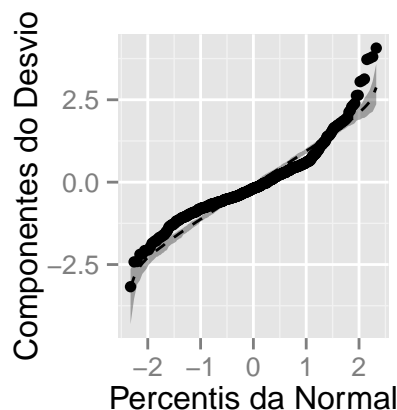


Figure 6: Gráfico Quantil-Quantil

No gráfico quantil-quantil (figura 6) podemos ver que muitos pontos saem da banda de confiança, indicando que a suposição de distribuição Gamma não é válida.

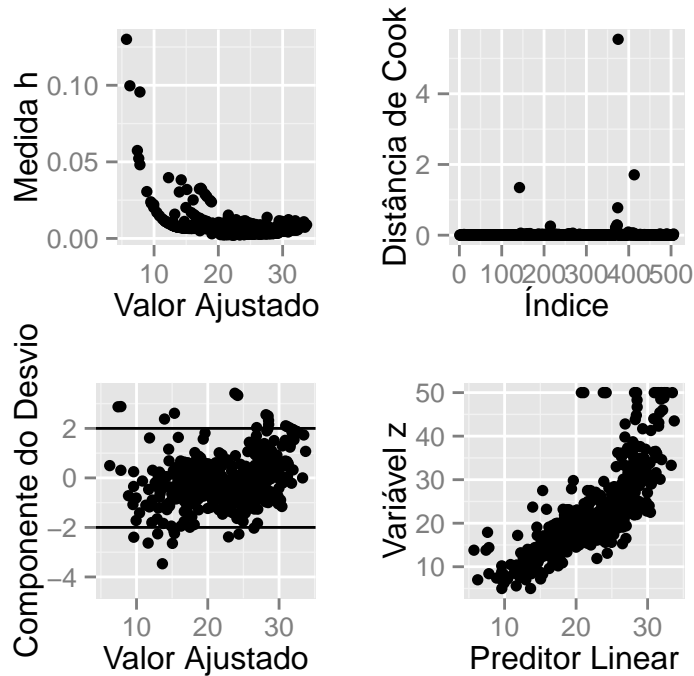


Figure 7: Gráficos de diagnóstico

Nos gráficos da figura 7 vemos não existe nenhum ponto com influência elevada. No gráfico do Valor ajustado pelo componente do desvio, vemos que são poucos os pontos que saem do intervalo  $[-2, 2]$  indicando que o ajuste está razoável. Além disso no gráfico de da variável z pelo preditor linear verificamos uma tendência linear indicando que a função de ligação escolhida é adequada.

## Ligação recíproca

O modelo que será ajustado é de regressão Gamma com ligação logarítmica. Ele é da forma:

$$y_i \sim \text{Gama}(\mu_i, \phi)$$

De forma que:

$$\mu_i = \frac{1}{x_i \beta}$$

## Seleção das variáveis

Como o número de variáveis explicativas é pequeno, vamos ajustar todos os modelos possíveis e escolher aquele que tiver o maior AIC. Os modelos estão apresentados na tabela abaixo.

```
todos.modelos <- fitall(y = dboston$medv, x = dboston[,1:4], method = "glm",
  family = Gamma(link = "inverse"))
```

variaveis	logL	AIC	BIC	n
(Intercept), dist	-1782	-3570	-3582	506

variaveis	logL	AIC	BIC	n
(Intercept), ptratio	-1727	-3460	-3473	506
(Intercept), dist, tax	-1723	-3454	-3471	506
(Intercept), tax	-1723	-3453	-3465	506
(Intercept), dist, ptratio	-1718	-3445	-3462	506
(Intercept), dist, tax, ptratio	-1687	-3384	-3405	506
(Intercept), tax, ptratio	-1687	-3382	-3399	506
(Intercept), lstat	-1515	-3035	-3048	506
(Intercept), tax, lstat	-1510	-3029	-3046	506
(Intercept), dist, lstat	-1500	-3009	-3026	506
(Intercept), dist, tax, lstat	-1488	-2986	-3007	506
(Intercept), ptratio, lstat	-1489	-2986	-3003	506
(Intercept), tax, ptratio, lstat	-1488	-2986	-3007	506
(Intercept), dist, ptratio, lstat	-1474	-2959	-2980	506
(Intercept), dist, tax, ptratio, lstat	-1469	-2949	-2975	506

O modelo com maior AIC é o modelo com todas as variáveis explicativas do banco de dados: **dist**, **tax**, **ptratio** e **lstat**. Mas pelo princípio de parcimônia vamos usar um modelo com menos variáveis que também tem o AIC alto, este modelo tem as seguintes variáveis: **dist**, **ptratio** e **lstat**.

Este modelo pode ser ajustado no R usando o seguinte comando:

```
modelo <- glm(medv ~ dist + ptratio + lstat, data = dboston,
              family = Gamma(link = "inverse"))
```

## Análise de Diagnóstico

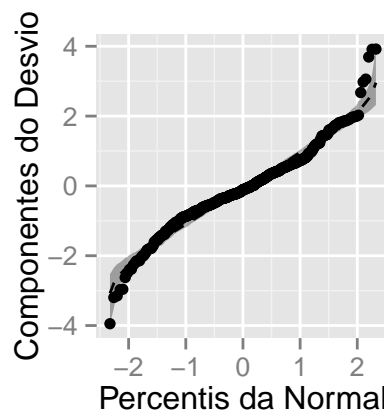


Figure 8: Gráfico Quantil-Quantil

No gráfico quantil-quantil (figura 8) podemos ver que muitos pontos saem da banda de confiança, indicando que a suposição de distribuição Gamma não é válida.



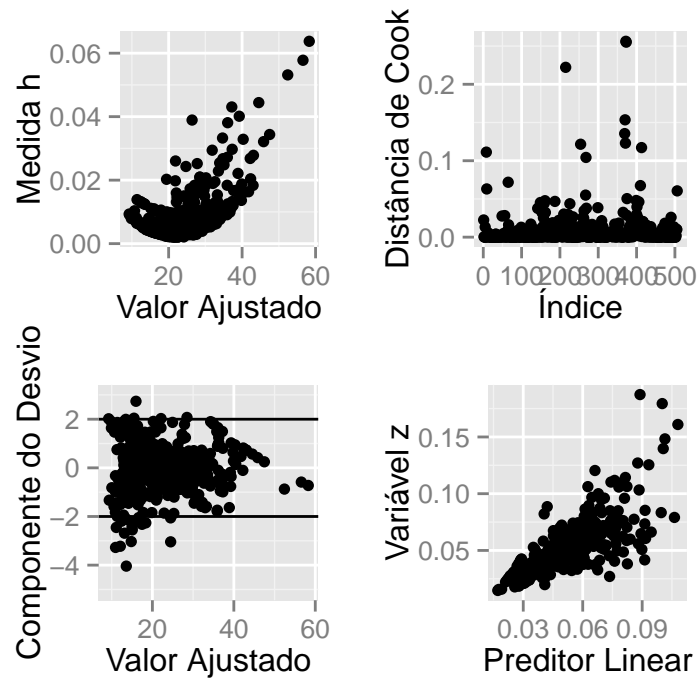


Figure 9: Gráficos de diagnóstico

Nos gráficos da figura 9 vemos não existe nenhum ponto com influência elevada. No gráfico do Valor ajustado pelo componente do desvio, vemos que são poucos os pontos que saem do intervalo  $[-2, 2]$  indicando que o ajuste está razoável. Além disso no gráfico de da variável  $z$  pelo preditor linear verificamos uma tendência linear indicando que a função de ligação escolhida é adequada.

## Modelo escolhido

Escolhemos como modelo final aquele com lição log, uma vez que