ME720 - Modelos Lineares Generalizados

Parte 10 - Teste de Hipóteses em MLGs

Profa. **Larissa Avila Matos**

Teste de hipóteses

Teste de hipóteses

A inferência para MLGs tem três abordagens usando a função de verossimilhança:

1. Teste da razão de verossimilhanças;
2. Teste de Wald;
3. Teste de Escore.

Focaremos em testes

$$H_0 : \beta = \beta_0 \quad \text{vs.} \quad H_1 : \beta \neq \beta_0.$$

Teste da razão de verossimilhanças

O teste da razão de verossimilhanças usa a função de verossimilhança através da razão de (1) seu valor L_0 em β_0, e (2) seu valor máximo L_1 nos valores de β permitindo que H_0 ou H_1 sejam verdadeiras.

A razão $\Lambda = \frac{L_0}{L_1} \leq 1$, uma vez que L_0 resulta da maximização em um valor restrito de β. A estatística do teste da razão de verossimilhanças é dada por

$$-2\log(\Lambda) = -2\log\left(\frac{L_0}{L_1}\right) = -2(\ell_0 - \ell_1),$$

onde ℓ_0 e ℓ_1 denotam as funções de log-verossimilhança maximizadas. Sob condições de regularidade, $-2\log(\Lambda)$ possui distribuição qui-quadrado quando $n \to \infty$, com $g.l = 1$ Então,

$$-2\log(\Lambda) \sim \chi^2_1, \quad n \to \infty.$$

O $p - valor$ é a probabilidade da qui-quadrado ser maior igual ao valor da estatística do teste observado, ou seja, $p - valor = \mathbb{P}(\chi^2_1 \geq \chi_{obs})$.

Este teste se estende diretamente a vários parâmetros. Por exemplo, para $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, considere $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$.

Então, L_1 é a função de verossimilhança calculada no valor de $\boldsymbol{\beta}$ pelo qual os dados seriam mais prováveis e L_0 é a função de verossimilhança calculada no valor de $\boldsymbol{\beta}$ para o qual os dados seriam mais prováveis quando $\boldsymbol{\beta} = \mathbf{0}$.

O grau de liberdade da distribuição da estatística do teste é igual à diferença nas dimensões dos espaços paramétricos em $H_0 \cup H_1$ e em H_0.

O teste também se estende à hipótese linear geral

$$H_0 : \boldsymbol{C\beta} = \mathbf{0},$$

uma vez que as restrições lineares implicam um novo modelo que é um caso especial do original.

Teste de Wald

Erros padrões obtidos a partir da inversa da matriz de informação dependem dos valores desconhecidos dos parâmetros.

Quando substituímos as estimativas irrestritas de MV (ou seja, não assumindo a hipótese nula), obtemos um erro padrão estimado (SE) de $\widehat{\beta}$.

Para $H_0 : \beta = \beta_0$ a estatística do teste usando esse erro padrão estimado não nulo,

$$z = \frac{\widehat{\beta} - \beta_0}{SE},$$

é chamada de estatística de Wald e segue aproximadamente uma distribuição normal padrão quando $\beta = \beta_0$.

Além disso, z^2 tem aproximadamente uma distribuição qui-quadrado com 1 grau de liberdade, $z^2 \sim \chi^2_1$.

Para vários parâmetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, para testar $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$, a estatística qui-quadrado de Wald é

$$\widehat{\boldsymbol{\beta}_0}' \left[\text{Var}(\widehat{\boldsymbol{\beta}_0}) \right]^{-1} \widehat{\boldsymbol{\beta}_0},$$

onde $\widehat{\boldsymbol{\beta}_0}$ é a estimativa de máxima verossimilhança irrestrita de $\boldsymbol{\beta}_0$ e $\text{Var}(\widehat{\boldsymbol{\beta}_0})$ é um bloco da matriz de covariância estimada irrestrita de $\widehat{\boldsymbol{\beta}}$.

Teste de Escore

O teste de Escore, usa a inclinação (ou seja, a função Escore) e a curvatura esperada da função de log-verossimilhança, avaliada no valor da hipótese nula β_0.

A estatística do teste é dada por

$$\frac{\left[\partial \ell(\beta)/\partial \beta_0\right]^2}{-\mathbb{E}\left[\partial \ell^2(\beta)/\partial \beta_0^2\right]} \sim \chi^2_1.$$

Para vários parâmetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, para testar $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$, a estatística qui-quadrado do teste de Escore é uma forma quadrática baseada no vetor de derivadas parciais da função de log-verossimilhança e na inversa da matriz de informação, ambas avaliadas nas estimativas sob H_0.

Função Desvio

Função Desvio ou Desvio

A função desvio é útil para termos uma idéia da adequabilidade do modelo (verificação da qualidade do ajuste).

Sem perda de generalidade, seja $\ell(\boldsymbol{\mu}; \mathbf{y}) \equiv \ell(\boldsymbol{\beta}, \phi; \mathbf{y})$ a log verossimilhança associada ao modelo em estudo (a log verossimilhança expressada em termos da média), então

$$\ell(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} \ell(\mu_i; \mathrm{y}_i),$$

em que $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j$.

Seja também $\ell(\mathbf{y};\mathbf{y})$ a log verossimilhança do modelo saturado ($n = p$), em que cada média é representada por ela mesma.

A função $\ell(\boldsymbol{\mu};\mathbf{y})$ é estimada por

$$\ell(\mathbf{y};\mathbf{y}) = \sum_{i=1}^{n} \ell(\mathrm{y}_i;\mathrm{y}_i),$$

ou seja, a estimativa de máxima verossimilhança de μ_i fica nesse caso dada por $\widehat{\mu_i} = \mathrm{y}_i$.

Quando $p < n$, denotamos a estimativa de $\ell(\boldsymbol{\mu};\mathbf{y})$ por $\ell(\widehat{\boldsymbol{\mu}};\mathbf{y})$.

O modelo saturado, explica toda variação pelo preditor linear do modelo.

Um modelo perfeito parece bom, mas o modelo saturado não é útil. Ele não suaviza os dados ou tem as vantagens de um modelo mais simples devido à sua parcimônia.

No entanto, muitas vezes serve como linha de base para comparação com outros modelos, como para verificar a bondade de ajuste.

A qualidade do ajuste de um MLG é avaliada através da função desvio

$$D^\star(\mathbf{y};\widehat{\boldsymbol{\mu}}) = 2\left[\ell(\mathbf{y};\mathbf{y}) - \ell(\widehat{\boldsymbol{\mu}};\mathbf{y})\right],$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros) avaliado na estimativa de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$.

Uma vez que $\ell(\widehat{\boldsymbol{\mu}};\mathbf{y}) \leq \ell(\mathbf{y};\mathbf{y})$, então $D^\star(\mathbf{y};\widehat{\boldsymbol{\mu}}) \geq 0$.

Um valor pequeno para a função desvio indica que, para um número menor de parâmetros, obtemos um ajuste tão bom quanto o ajuste com o modelo saturado.

Denotando por $\widehat{\theta}_i = \theta_i(\widehat{\mu}_i)$ e $\widetilde{\theta}_i = \theta_i(\widetilde{\mu}_i)$ as estimativas de máxima verossimilhança de θ_i para os modelos com p parâmetros ($p < n$) e saturado ($p = n$), respectivamente, temos que a função $D^\star(\mathbf{y}; \boldsymbol{\mu})$ fica, alternativamente, dada por

$$\begin{aligned} D^\star(\mathbf{y}; \widehat{\boldsymbol{\mu}}) &= 2\sum_{i=1}^{n}\left[\frac{\mathrm{y}_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)}{a(\phi)}\right] - 2\sum_{i=1}^{n}\left[\frac{\mathrm{y}_i\widehat{\theta}_i - b(\widehat{\theta}_i)}{a(\phi)}\right] \\ &= 2\sum_{i=1}^{n}\left[\frac{\mathrm{y}_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)}{a(\phi)}\right]. \end{aligned}$$

Usualmente $a(\phi) = \frac{\phi}{w_i}$, então

$$\frac{D(\mathbf{y}; \widehat{\boldsymbol{\mu}})}{\phi} = 2\sum_{i=1}^{n} w_i \left[\mathrm{y}_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i) \right],$$

$\frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}})}{\phi}$ é chamado de desvio escalonado. A estatística $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})$ é chamada de desvio.

Ou seja, nesse caso $D^\star(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y};\widehat{\boldsymbol{\mu}})}{\phi}$ é o desvio escalonado.

- **Obs.**: O desvio no R sai com o nome de deviance após o ajuste do modelo e o número de graus de liberdade correspondente é dado por $n - p$.

Razão de verossimilhanças e Diferença de desvios

Métodos para comparar desvios generalizam métodos para modelos lineares normais que comparam somas de quadrados de resíduos.

Quando $\phi = 1$, como para um modelo de Poisson ou binomial, a função desvio é igual a

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = -2\left[\ell(\widehat{\boldsymbol{\mu}}; \mathbf{y}) - \ell(\mathbf{y}; \mathbf{y})\right].$$

Considere dois modelos encaixados, M_0 com parâmetro p_0 e valores ajustados $\widehat{\boldsymbol{\mu}_0}$ e M_1 com parâmetro p_1 e valores ajustados $\widehat{\boldsymbol{\mu}_1}$, sendo M_0 um caso especial de M1.

Obs: Os modelos M_0 e M_1 serem encaixados significa que o modelo M_0 é um caso particular de M_1. Por exemplo, M_0 é um modelo linear e M_1 é um modelo quadrático.

Como o espaço paramétrico de M_0 está contido no espaço de M_1, temos que

$$\ell(\widehat{\boldsymbol{\mu}_0}; \mathbf{y}) \leq \ell(\widehat{\boldsymbol{\mu}_1}; \mathbf{y}).$$

Como $L(\mathbf{y}; \mathbf{y})$ é idêntico para cada modelo,

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}_1}) \leq D(\mathbf{y}; \widehat{\boldsymbol{\mu}_0}).$$

Modelos mais simples têm desvios maiores.

Supondo que o modelo M_1 seja válido, o teste da razão de verossimilhança, com H_0: o modelo M_0 é preferível ao modelo M_1, usa a estatística do teste

$$\begin{aligned} -2\left[\ell(\widehat{\boldsymbol{\mu}_0};\mathbf{y}) - \ell(\widehat{\boldsymbol{\mu}_1};\mathbf{y})\right] &= -2\left[\ell(\widehat{\boldsymbol{\mu}_0};\mathbf{y}) - \ell(\mathbf{y};\mathbf{y})\right] - \left\{-2\left[\ell(\widehat{\boldsymbol{\mu}_1};\mathbf{y}) - \ell(\mathbf{y};\mathbf{y})\right]\right\} \\ &= D(\mathbf{y};\widehat{\boldsymbol{\mu}_0}) - D(\mathbf{y};\widehat{\boldsymbol{\mu}_1}), \end{aligned}$$

quando $\phi = 1$. O valor dessa estatística é grande quando M_0 é mal ajustado comparado com M_1.

Na expressão da função desvio, uma vez que os termos que envolvem o modelo saturado cancelam,

$$D(\mathbf{y};\widehat{\boldsymbol{\mu}_0}) - D(\mathbf{y};\widehat{\boldsymbol{\mu}_1}) = 2\sum_i \omega_i \left[\mathrm{y}_i\left(\widehat{\theta_{1i}} - \widehat{\theta_{0i}}\right) - b\left(\widehat{\theta_{1i}}\right) + b\left(\widehat{\theta_{0i}}\right)\right],$$

também tem a forma do desvio.

Sob condições de regularidade, onde a estatística da razão de verossimilhança têm distribuição qui-quadrado para amostra grande, essa diferença tem aproximadamente uma distribuição qui-quadrado, com $p_1 - p_0$ graus de liberdade.

Denotamos a estatística da razão de verossimilhança para comparar modelos encaixados por $G^2(M_0|M_1)$.

Exemplo: Para um modelo Poisson log-linear com intercepto, a diferença dos desvios usa as contagens observadas e os dois conjuntos de valores ajustados na forma

Sob condições de regularidade, onde a estatística da razão de verossimilhança têm distribuição qui-quadrado para amostra grande, essa diferença tem aproximadamente uma distribuição qui-quadrado, com $p_1 - p_0$ graus de liberdade.

Denotamos a estatística da razão de verossimilhança para comparar modelos encaixados por $G^2(M_0|M_1)$.

Exemplo: Para um modelo Poisson log-linear com intercepto, a diferença dos desvios usa as contagens observadas e os dois conjuntos de valores ajustados na forma

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}_0}) - D(\mathbf{y}; \widehat{\boldsymbol{\mu}_1}) = 2 \sum_i \mathrm{y}_i \log \left(\frac{\widehat{\mu_{1i}}}{\widehat{\mu_{0i}}} \right).$$

Teste Escore e Estatística de Pearson para comparação de modelos

Para MLGs com função de variância $\text{Var}(\text{y}_i) = V(\mu_i)$ com $\phi = 1$, a estatística de escore para comparar o modelo escolhido com o modelo saturado é

$$X^2 = \sum_i \frac{(\text{y}_i - \widehat{\mu_i})^2}{V(\widehat{\mu_i})}.$$

Para y_i Poisson, onde $V(\widehat{\mu_i}) = \widehat{\mu_i}$, temos que a estatística de escore é dada por

$$X^2 = \sum_i \frac{(\text{observado} - \text{ajustado})^2}{\text{ajustado}}.$$

Esta estatística é conhecida como estatística qui-quadrado de Pearson, pois Karl Pearson a introduziu em 1900 para testar várias hipóteses usando a distribuição qui-quadrado, com a hipótese nula de independência em uma tabela de contingência.

A estatística generalizada de Pearson é uma alternativa para a função desvio em testes de certos MLGs.

Para dois modelos encaixados, a estatística generalizada de Pearson para comparar esses modelos é dada por

$$X^2(M_0|M_1) = \sum_i \frac{\left(\widehat{\mu_{1i}} - \widehat{\mu_{0i}}\right)^2}{V(\widehat{\mu_{0i}})}.$$

Esta é uma aproximação quadrática para $G^2(M_0|M_1)$, com o mesmo comportamento assintótico sob H_0. No entanto, essa não é a estatística de escore para comparar os modelos.

Resíduos e valores ajustados

Os resíduos e os valores ajustados são assintoticamente não correlacionados.

A análise dos resíduos nos ajuda a identificar onde o ajuste de um MLG é ruim ou onde ocorrem observações discrepantes. Como nos modelos lineares normais, gostaríamos de explorar a decomposição

$$\boldsymbol{Y} = \widehat{\boldsymbol{\mu}} + (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}), \quad \text{(isto é, dados = ajuste + resíduos)}.$$

Com MLGs, no entanto, $\widehat{\boldsymbol{\mu}}$ e $\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}$ não são ortogonais.

A definição de um resíduo studentizado para os MLGs pode ser feita analogamente à regressão normal linear. No entanto, não necessariamente as propriedades continuam valendo.

Assim, torna-se importante a definição de outros tipos de resíduo cujas propriedades sejam conhecidas ou pelo menos estejam mais próximas das propriedades dos resíduos studentizados.

Podemos obter a matriz de covariância assintótica para os resíduos ($\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{\mu}}$).

Vimos que $\boldsymbol{W} = \text{diag}\left(\frac{(\partial\mu_i/\partial\eta_i)^2}{\text{Var}(Y_i)}\right)$ e $\boldsymbol{D} = \text{diag}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)$, então podemos expressar a matriz diagonal $\boldsymbol{V} = \text{Var}(\boldsymbol{Y})$ como

$$\boldsymbol{V} = \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D}.$$

Para n grande, se $\widehat{\boldsymbol{\mu}}$ é aproximadamente não correlacionado com $\boldsymbol{R}$, então

$$\boldsymbol{V} \approx \text{Var}\left(\widehat{\boldsymbol{\mu}}\right) + \text{Var}\left(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}\right).$$

Vimos também que

$$\text{Var}(\widehat{\boldsymbol{\mu}}) \approx \boldsymbol{D}'\text{Var}(\widehat{\boldsymbol{\eta}})\boldsymbol{D} \approx \boldsymbol{D}\boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{D}.$$

Portanto, usando $\boldsymbol{V}^{1/2} = \boldsymbol{D}\boldsymbol{W}^{-1/2}$, temos que

$$\begin{aligned}
\text{Var}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) &\approx \boldsymbol{V} - \text{Var}\,(\widehat{\boldsymbol{\mu}}) \approx \boldsymbol{D}\boldsymbol{W}^{-1}\boldsymbol{D} - \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D} \\
&\approx \boldsymbol{D}\boldsymbol{W}^{-1/2}\left[\boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1/2}\right]\boldsymbol{W}^{-1/2}\boldsymbol{D} \\
&\approx \boldsymbol{V}^{1/2}[\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{W}}]\boldsymbol{V}^{1/2},
\end{aligned}$$

onde

$$\boldsymbol{H}_{\boldsymbol{W}} = \left[\boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2}\right].$$

Podemos verificar que $\boldsymbol{H}_{\boldsymbol{W}}$ é uma matriz de projeção mostrando que é simétrica e idempotente. McCullagh e Nelder (1989) observaram que $\boldsymbol{H}_{\boldsymbol{W}}$ é aproximadamente uma matriz chapéu para unidades padronizadas de Y, com

$$\boldsymbol{H}_{\boldsymbol{W}}\boldsymbol{V}^{-1/2}(\boldsymbol{Y} - \boldsymbol{\mu}) \approx \boldsymbol{H}_{\boldsymbol{W}}\boldsymbol{V}^{-1/2}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}).$$

Para mais detalhes, olhar o livro texto.

Resíduo de Pearson

Para um modelo específico com função de variância $V(\mu)$ o resíduo de Pearson para observação y_i e seu valor ajustado é

$$e_i = \frac{\mathrm{y}_i - \widehat{\mu_i}}{\sqrt{V(\widehat{\mu_i})}}.$$

A soma de seus valores ao quadrado resultam na estatística generalizada de Pearson.

Resíduo desvio

Um outro resíduo é o resíduo desvio.

O resíduo desvio é dado por

$$\sqrt{d_i^\star}\ \text{sinal}(\text{y}_i - \widehat{\mu}_i),$$

onde a função desvio é dada por $D^\star(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = \sum_i d_i^\star$, ou seja,

$$d_i^\star = 2\omega_i \left[\text{y}_i \left(\widetilde{\theta}_i - \widehat{\theta}_i \right) - b\left(\widetilde{\theta}_i \right) + b\left(\widehat{\theta}_i \right) \right].$$

A soma dos quadrados dos resíduos desvio é igual ao desvio.

Para julgar quando um resíduo é “grande”, é útil ter valores residuais que, quando o modelo é válido, possui média 0 e variância 1.

No entanto, os resíduos de Pearson e de desvio tendem a ter variação menor que 1, pois eles comparam y_i com a média ajustada $\widehat{\mu_i}$ em vez da média verdadeira μ_i.

O resíduo padronizado divide cada resíduo $(\mathrm{y}_i - \widehat{\mu_i})$ pelo seu erro padrão.

Vimos que $\mathrm{Var}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) \approx \boldsymbol{V}^{1/2}[\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{W}}]\boldsymbol{V}^{1/2}$, ou seja,

$$\mathrm{Var}(Y_i - \widehat{\mu_i}) \approx V(\mu_i)(1 - h_{ii}),$$

onde h_{ii} é o elemento da diagonal da matriz chapéu generalizada $\boldsymbol{H}_{\boldsymbol{W}}$ para a observação i, sua alavanca.

Resíduo padronizado

Seja $\widehat{h_{ii}}$ a estimativa de h_{ii}.

Então, padronizando o resíduo de Pearson dividindo pelos seus erros estimados (SE), temos que

$$r_i = \frac{\mathrm{y}_i - \widehat{\mu_i}}{\sqrt{V(\widehat{\mu_i})(1 - \widehat{h_{ii}})}} = \frac{e_i}{\sqrt{1 - \widehat{h_{ii}}}},$$

onde r_i é o resíduo padronizado para a observação i.

Williams (1984) mostra através de estudos de Monte Carlo que a distribuição de r_i é em geral assimétrica, mesmo para grandes amostras.

Da mesma forma, os resíduos de desvio têm versões padronizadas.

Eles são mais úteis para casos assintóticos de pequena dispersão, como para médias de Poisson relativamente grandes e índices binomiais relativamente grandes. Nesses casos, sua distribuição baseada em modelo é aproximadamente padrão normal.

Resíduo desvio padronizado

Os resíduos mais utilizados em modelos lineares generalizados são definidos a partir dos componentes da função desvio. A versão padronizada (ver McCullagh, 1987; Davison e Gigli, 1989) é dada por

$$t_{d_i} = \frac{\sqrt{d_i^\star}}{\sqrt{1-\widehat{h_{ii}}}}\ \text{sinal}(\mathrm{y}_i - \widehat{\mu}_i),$$

onde $d_i^\star = 2\omega_i \left[\mathrm{y}_i\left(\widetilde{\theta}_i - \widehat{\theta}_i\right) - b\left(\widetilde{\theta}_i\right) + b\left(\widehat{\theta}_i\right)\right]$.

Williams (1984) verificou através de simulações que a distribuição de t_{d_i} tende a estar mais próxima da normalidade do que as distribuições de outros resíduos (veja Paula (2013)).

Pode acontecer de que o modelo esteja bem ajustado e, mesmo assim, a distribuição do resíduo desvio padronizado pode nãoo ser aproximadamente normal.

Ainda assim podemos construir um gráfico de quantil quantil com envelopes simulando a partir do modelo de interesse ao invés da distribuição normal.

Já vimos como construir um gráfico de quantil quantil para o modelos lineares normais.

Casos particulares

- Poisson:

$$t_{d_i} = \frac{\sqrt{2\left[\mathrm{y}_i \log\left(\frac{\mathrm{y}_i}{\widehat{\mu}_i}\right) - \mathrm{y}_i + \widehat{\mu}_i\right]}}{\sqrt{1-\widehat{h_{ii}}}}\ \mathrm{sinal}(\mathrm{y}_i - \widehat{\mu}_i)$$

$$= \frac{\sqrt{2\widehat{\mu}_i}}{\sqrt{1-\widehat{h_{ii}}}}\ \mathrm{sinal}(\mathrm{y}_i - \widehat{\mu}_i) \quad \text{se } \mathrm{y}_i = 0.$$

- Normal:

$$t_{d_i} = \frac{(\mathrm{y}_i - \widehat{\mu}_i)}{\sqrt{\widehat{\sigma^2}(1-\widehat{h_{ii}})}}\ \mathrm{sinal}(\mathrm{y}_i - \widehat{\mu}_i).$$

Para detectar a falta de ajuste de um modelo, qualquer tipo particular de resíduo pode ser graficado em relação aos valores ajustados do componente em $\widehat{\boldsymbol{\mu}}$ e em relação a cada variável explicativa.

Assim como no modelo linear, o ajuste pode ser bem diferente quando excluímos uma observação que possui um grande resíduo padronizado e uma grande alavancagem.

Como vimos, as alavancas estimadas caem entre 0 e 1 e somam p.

Diferentemente dos modelos lineares comuns, a matriz chapéu generalizada depende do ajuste e da matriz do modelo, e os pontos que possuem valores extremos para as variáveis explicativas não precisam ter alta alavancagem estimada.

Para medir a influência, uma distância análoga da distância de Cook usa os resíduos padronizados e as alavancas estimadas, dada por

$$r_i^2 \left[\frac{\widehat{h_{ii}}}{p(1-\widehat{h_{ii}})} \right].$$

Outras estatísticas de comparação de modelos

O teste é apropriado na comparação de modelos encaixados.

Além disso, ele não leva em consideração (diretamente) o número de parâmetros do modelo (somente na distribução da estatística).

Existem várias alternativas, em termos de estatística para comparar modelos, que "penalizam" a verossimilhança em relação ao número de parâmetros, tamanho da amostra entre outros fatores.

Veremos o AIC e o BIC.

AIC e BIC

O AIC (*Akaike information criterion*) e BIC (*Bayesian information criterion*) são dados, respectivamente, por:

$$\text{AIC} = -2(\ell(\widehat{\boldsymbol{\beta}_M}) - k) = -2\ell(\widehat{\boldsymbol{\beta}_M}) + 2k, \quad \text{e}$$

$$\text{BIC} = -2(\ell(\widehat{\boldsymbol{\beta}_M}) - k\log(n)) = -2\ell(\widehat{\boldsymbol{\beta}_M}) + 2k\log(n);$$

em que $\ell(\widehat{\boldsymbol{\beta}_M})$ denota a log-verossimilhança do modelo avaliada na estimativa MV, k é o número de parâmetros em M e n é o número de de observações.

Portanto, o modelo que apresentar os menores valores, será o modelo "melhor ajustado" aos dados.

Seleção de variáveis explicativas

A seleção de modelos para MLGs enfrenta os mesmos problemas dos modelos lineares comuns.

O processo de seleção se torna mais difícil à medida que aumentamos o número de variáveis explicativas, devido ao rápido aumento de possíveis efeitos e interações.

O processo de seleção tem dois objetivos concorrentes:

- O modelo deve ser suficientemente complexo para ajustar os dados.
- Por outro lado, deve suavizar em vez de ajustar demais (*overfit*) os dados e, idealmente, ser relativamente simples de interpretar.

A maioria dos estudos de pesquisa é projetada para responder a certas perguntas. Essas perguntas orientam a escolha dos termos do modelo.

As análises confirmatórias usam um conjunto restrito de modelos.

Por exemplo, uma hipótese de estudo sobre um efeito pode ser testada comparando modelos com e sem esse efeito.

Para estudos exploratórios e não confirmatórios, uma pesquisa entre possíveis modelos pode fornecer pistas sobre a estrutura dos efeitos e levantar questões para futuras pesquisas.

Em qualquer um dos casos, é útil primeiro estudar o efeito marginal de cada preditor por si só, com estatística descritiva e uma matriz de gráficos de dispersão, para ter uma ideia desses efeitos.

Procedimento *Stepwise*

Seleção *Forward* e eliminação *Backward*:

Com p variáveis explicativas, o número de modelos possíveis é 2^p, pois cada variável está ou não no modelo escolhido.

A melhor seleção de subconjunto identifica o modelo com melhor desempenho de acordo com um critério, como p.e. maximizar o valor R^2 ajustado.

Isso é computacionalmente intensivo quando p é grande.

Métodos alternativos podem pesquisar entre todos os modelos.

Métodos *Forward* e *Backward*

Esse métodos são iguais para os modelos lineares normais, onde a diferença para os MLGs é que as covariáveis estão no preditor linear.

Forward: esse método adiciona termos sequencialmente.

Backward: esse método começa com um modelo complexo e remove os termos sequencialmente.

Para mais detalhes ver o livro texto e Notas de aula do Prof. Gilberto de Paula.

Exemplo Livro

O que afeta o preço de venda de uma casa?

Os dados a seguir mostram as observações sobre as vendas de casas recentes em Gainesville, Flórida. O arquivo de dados para as 100 vendas de casas estão no site do livro texto.

As variáveis listadas são:

- o preço de venda (em milhares de dólares),
- o tamanho da casa (em pés quadrados),
- a conta anual do imposto predial (em dólares),
- o número de quartos,
- o número de banheiros, e
- se a casa é nova.

Como essas 100 observações são de uma única cidade, não podemos usá-las para fazer inferências sobre as relações em geral.

Mas, para fins ilustrativos, os tratamos como uma amostra aleatória de uma população conceitual de vendas de casas nesse mercado e analisamos como o preço de venda parece se relacionar com essas características.

A seguir são apresentados os dados de 5 casas do conjunto de dados e algumas análises iniciais.

```r
Houses <- as.data.frame(read.table('Houses.txt', header = T))
Houses[1:5,]
```

```
  case taxes beds baths new price size
1    1  3104    4     2   0 279.9 2048
2    2  1173    2     1   0 146.5  912
3    3  3076    4     2   0 237.7 1654
4    4  1608    3     2   0 200.0 2068
5    5  1454    3     3   0 159.9 1477
```

```r
str(Houses)
```

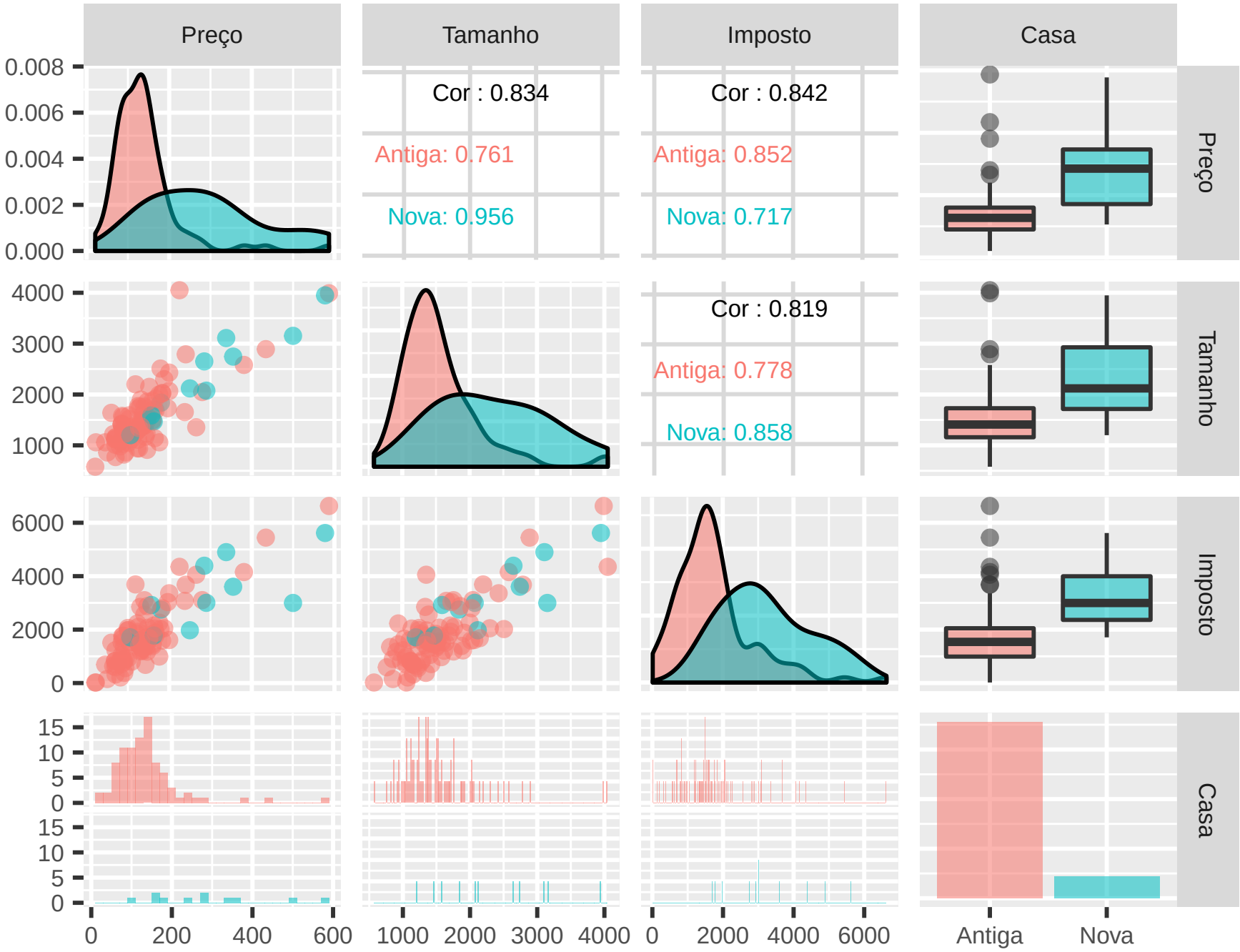
```
'data.frame':   100 obs. of  7 variables:
 $ case : int  1 2 3 4 5 6 7 8 9 10 ...
 $ taxes: int  3104 1173 3076 1608 1454 2997 4054 3002 6627 320 ...
 $ beds : int  4 2 4 3 3 3 3 3 5 3 ...
 $ baths: int  2 1 2 2 3 2 2 2 4 2 ...
 $ new  : int  0 0 0 0 0 1 0 1 0 0 ...
 $ price: num  280 146 238 200 160 ...
 $ size : int  2048 912 1654 2068 1477 3153 1355 2075 3990 1160 ...
```

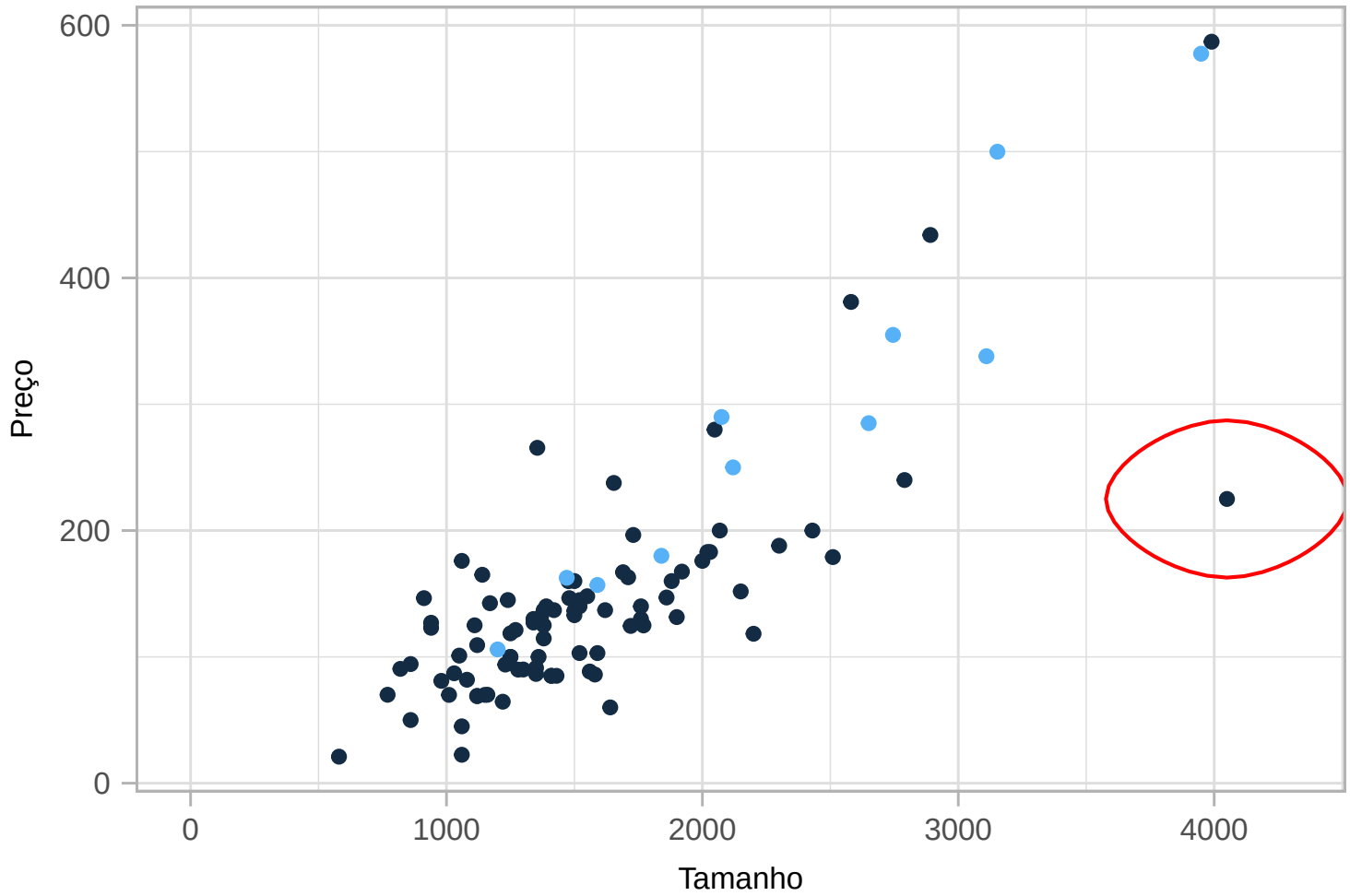
```r
cbind(mean(price), sd(price), mean(size), sd(size))
```

```
        [,1]     [,2]    [,3]     [,4]
[1,] 155.331 101.2622 1629.28 666.9417
```

```r
table(new)
```

```
new
 0  1
89 11
```


Preço
Tamanho
Imposto
Casa
Cor : 0.834
Antiga: 0.761
Nova: 0.956
Cor : 0.842
Antiga: 0.852
Nova: 0.717
Cor : 0.819
Antiga: 0.778
Nova: 0.858
0.008
0.006
0.004
0.002
0.000
4000
3000
2000
1000
6000
4000
2000
0
15
10
5
0
0
200
400
600
1000
2000
3000
4000
0
2000
4000
6000
Antiga
Nova



Temos aproximadamente uma tendência linear crescente para o preço de venda em função do tamanho. Uma exceção é um preço de venda relativamente baixo para uma residência muito grande que não era nova (observação 64 no campo dos dados). Apenas 11 casas da amostra eram novas, portanto o impacto dessa variável é pouco claro.

Ajuste linear para esse conjunto de dados pode ser encontrado no livro texto.

Em seguida, ajustamos o modelo gamma

```
fit.gamma <- glm(price ~ size + new + beds + size:new + size:beds, family = Gamma(link = identity), data=Houses)
summary(fit.gamma)
```

```
Call:
glm(formula = price ~ size + new + beds + size:new + size:beds,
    family = Gamma(link = identity), data = Houses)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.10123  -0.25618  -0.02687   0.12514   0.94699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.37590   48.59784   0.913   0.3635
size          0.07398    0.04000   1.850   0.0675 .
new         -60.02905   65.76551  -0.913   0.3637
beds        -22.71311   17.63124  -1.288   0.2008
size:new      0.05383    0.03758   1.432   0.1553
size:beds     0.01000    0.01256   0.796   0.4279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1095153)

    Null deviance: 31.940  on 99  degrees of freedom
Residual deviance: 10.263  on 94  degrees of freedom
AIC: 1049

Number of Fisher Scoring iterations: 10
```

```r
fit.gamma <- glm(price ~ size + new + beds + size:new + size:beds, family = Gamma(link = inverse), data=Houses)
summary(fit.gamma)
```

```
Call:
glm(formula = price ~ size + new + beds + size:new + size:beds,
    family = Gamma(link = inverse), data = Houses)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.26381  -0.20322  -0.02222   0.17969   0.90695

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.538e-02  2.377e-03   6.468 4.43e-09 ***
size        -3.456e-06  8.141e-07  -4.245 5.13e-05 ***
new         -3.663e-03  1.416e-03  -2.587   0.0112 *
beds        -1.253e-03  7.673e-04  -1.632   0.1059
size:new     7.963e-07  4.521e-07   1.761   0.0814 .
size:beds    3.318e-07  2.416e-07   1.373   0.1729
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1241193)

    Null deviance: 31.940  on 99  degrees of freedom
Residual deviance: 12.863  on 94  degrees of freedom
AIC: 1072

Number of Fisher Scoring iterations: 16
```

```r
fit.g1 <- glm(price ~ size+new+baths+beds, family=Gamma(link=identity),data=Houses)
fit.g2 <- glm(price~(size+new+baths+beds)^2,family=Gamma(link=identity),data=Houses)
anova(fit.g1, fit.g2, test="F")
```

```
Analysis of Deviance Table

Model 1: price ~ size + new + baths + beds
Model 2: price ~ (size + new + baths + beds)^2
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        95    10.4417
2        89     9.8728  6  0.56894 0.8438 0.5396
```

```r
summary(fit.g1)
```

```
Call:
glm(formula = price ~ size + new + baths + beds, family = Gamma(link = identity),
    data = Houses)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.03363  -0.25134  -0.02568   0.11996   0.90519

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.00267   18.35742  -0.109   0.9134
size          0.10948    0.01392   7.866 5.77e-12 ***
new          33.05951   24.13653   1.370   0.1740
baths        10.29246    9.18244   1.121   0.2652
beds        -15.35424    8.38910  -1.830   0.0703 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1119279)

    Null deviance: 31.940  on 99  degrees of freedom
Residual deviance: 10.442  on 95  degrees of freedom
AIC: 1048.8

Number of Fisher Scoring iterations: 8
```

```r
summary(fit.g2)
```

```
Call:
glm(formula = price ~ (size + new + baths + beds)^2, family = Gamma(link = identity),
    data = Houses)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95965  -0.25435  -0.00945   0.10162   0.91413

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.198133  56.656449   1.257   0.2122
size          0.103301   0.057271   1.804   0.0747 .
new          -9.488648 187.470596  -0.051   0.9597
baths        -9.032257  43.400382  -0.208   0.8356
beds        -53.949807  29.198846  -1.848   0.0680 .
size:new      0.085562   0.055851   1.532   0.1291
size:baths   -0.017403   0.020726  -0.840   0.4033
size:beds     0.009647   0.019453   0.496   0.6212
new:baths   -70.241353 109.571186  -0.641   0.5231
new:beds     14.886093  55.853026   0.267   0.7905
baths:beds   16.124186  15.991339   1.008   0.3160
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1123827)

    Null deviance: 31.9401  on 99  degrees of freedom
Residual deviance:  9.8728  on 89  degrees of freedom
AIC: 1055.1

Number of Fisher Scoring iterations: 10
```

```r
summary(glm(price ~ size+new+size:new, family=Gamma(link=identity),data=Houses))
```

```
Call:
glm(formula = price ~ size + new + size:new, family = Gamma(link = identity),
    data = Houses)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.14747  -0.23680  -0.03117   0.09924   0.90883

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.45219   12.97384  -0.574   0.5670
size          0.09446    0.01005   9.396 2.95e-15 ***
new         -77.90333   64.58272  -1.206   0.2307
size:new      0.06492    0.03670   1.769   0.0801 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1102068)

    Null deviance: 31.940  on 99  degrees of freedom
Residual deviance: 10.563  on 96  degrees of freedom
AIC: 1047.9

Number of Fisher Scoring iterations: 6
```

```r
summary(lm(price ~ size + new + size:new, data=Houses))
```

```
Call:
lm(formula = price ~ size + new + size:new, data = Houses)

Residuals:
    Min      1Q  Median      3Q     Max
-175.75  -28.98   -6.26   14.69  192.52

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.227808  15.521110  -1.432  0.15536
size          0.104438   0.009424  11.082  < 2e-16 ***
new         -78.527502  51.007642  -1.540  0.12697
size:new      0.061916   0.021686   2.855  0.00527 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52 on 96 degrees of freedom
Multiple R-squared:  0.7443,    Adjusted R-squared:  0.7363
F-statistic: 93.15 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
AIC(lm(price ~ size + new + size:new, data=Houses))
```

```
[1] 1079.947
```

```
AIC(lm(price ~ size +new +beds +baths +size:new +
         size:beds +new:baths, data=Houses))
```

```
[1] 1070.565
```

Referência

- Paula, G.A. (2013). Modelos de Regressão com Apoio Computacional.
- Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley series in probability and statistics.