

Exercício 3

Descrição

No arquivo **imoveis.dat** são apresentados dados relativos a uma amostra de 27 imóveis. Na ordem são apresentados os valores das seguintes variáveis:

- imposto do imóvel (em 100 USD)
- área do terreno (em 1000 pés quadrados)
- área construída (em 1000 pés quadrados)
- idade da residência (em anos)
- preço de venda do imóvel (em 1000 USD)

Enunciado

- Ajuste um modelo normal linear do preço de venda contra as demais variáveis explicativas.
- Use o método AIC para selecionar as variáveis explicativas. Faça uma análise de diagnóstico com o modelo selecionado.
- Interprete os coeficientes estimados.
- Seja $y(z)$ o valor do preço de venda de um imóvel que não está na amostra com os valores das variáveis explicativas do modelo final representados por z . Como fica a estimativa intervalar de coeficiente $(1 - \alpha)$, $0 < \alpha < 1$, para $y(z)$? Alguma restrição para os valores de z ?

Leitura dos dados

Para ler os dados no R fazemos:

```
imoveis <- data.frame(  
  scan("dados/imoveis.dat", list(imposto=0, areat=0, areac=0, idade=0, preco=0)))
```

imposto	areat	areac	idade	preco
4.9	3.5	1.00	42	26
5.0	3.5	1.50	62	30
4.5	2.3	1.18	40	28
4.6	4.0	1.23	54	26
5.1	4.5	1.12	42	30
3.9	4.5	0.99	56	30

Análise descritiva

```
df <- gather(imoveis, key = var, value = value, -preco)
ggplot(df, aes(x=value, y=preco)) + geom_point() + stat_smooth(method = "lm", se = F) +
  facet_wrap(~var, scales = "free") + xlab("Valor da variável") + ylab("Preço")
```

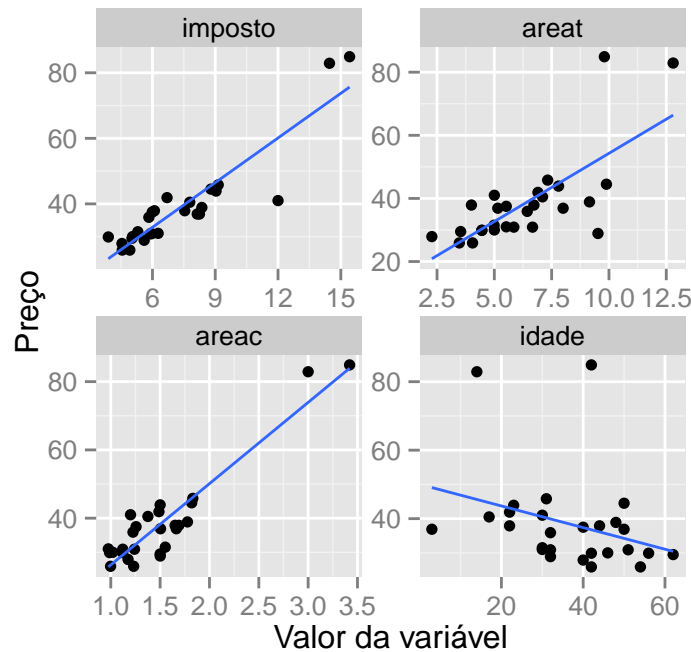


Figure 1: Gráfico de dispersão da variável resposta ‘Preço’ por todas as variáveis explicativas.

Vemos na figura 1 que todas as variáveis apresentam relação linear com a variável resposta. Quanto maior o imposto ou área de terreno ou área construída, maior a preço do imóvel. E quanto maior a idade do imóvel menor o preço do imóvel.

Modelo

Para selecionar as variáveis explicativas vamos usar o método AIC. Neste caso como o número de variáveis explicativas é baixo, vamos ajustar todos os modelos possíveis e depois escolher aquele com o maior AIC.

Usando o código abaixo é possível ajustar todos os modelos com as variáveis do banco de dados.

```
library(meifly)
todos.modelos <- fitall(imoveis$preco, imoveis[, -5])
```

variaveis	logL	AIC	BIC	R2	adjR2	n
(Intercept), idade	-108	-223	-226	0.10	0.06	27
(Intercept), areat, idade	-99	-207	-212	0.53	0.49	27
(Intercept), areat	-100	-205	-209	0.53	0.51	27
(Intercept), imposto, idade	-85	-178	-183	0.84	0.83	27

variaveis	logL	AIC	BIC	R2	adjR2	n
(Intercept), imposto, areat, idade	-83	-176	-183	0.86	0.84	27
(Intercept), imposto	-85	-176	-180	0.84	0.83	27
(Intercept), imposto, areat	-84	-175	-180	0.86	0.84	27
(Intercept), areac	-83	-172	-176	0.86	0.86	27
(Intercept), areat, areac	-82	-172	-177	0.87	0.86	27
(Intercept), areat, areac, idade	-80	-170	-177	0.89	0.87	27
(Intercept), areac, idade	-81	-169	-174	0.88	0.87	27
(Intercept), imposto, areat, areac, idade	-73	-159	-167	0.93	0.92	27
(Intercept), imposto, areat, areac	-74	-158	-164	0.93	0.92	27
(Intercept), imposto, areac, idade	-74	-157	-164	0.93	0.92	27
(Intercept), imposto, areac	-74	-156	-161	0.93	0.92	27

Vemos pela tabela acima que o modelo com maior AIC é aquele com as variáveis ‘imposto’ e ‘areac’ (área construída). Coincidentemente este modelo também apresenta o maior R^2 ajustado.

Portanto o modelo linear normal escolhido pelo método AIC é da forma:

$$(\text{preço})_i = \beta_0 + \beta_1(\text{imposto})_i + \beta_2(\text{areac})_i + \epsilon_i$$

Com $\epsilon_i \sim \text{Normal}(0, \sigma^2)$.

O ajuste foi feito no R com o comando a seguir:

```
modelo <- lm(preco ~ imposto + areac, data = imoveis)
```

E as estimativas estão apresentadas na tabela abaixo:

	Estimativa	Erro Padrão	valor z	Pr(> t)
β_0	0.7903	2.2794	0.35	0.7318
β_1	2.2971	0.4890	4.70	0.0001
β_2	13.9333	2.5244	5.52	0.0000

Como neste modelo o intercepto (β_0) não é significativo (valor-p = 0.73), vamos reajustar um modelo sem este parâmetro. O ajuste pode ser feito no R com o comando a seguir:

```
modelo <- lm(preco ~ 0 + imposto + areac, data = imoveis)
```

As novas estimativas estão abaixo:

	Estimativa	Erro Padrão	valor z	Pr(> t)
β_1	2.3242	0.4741	4.90	0.0000
β_2	14.2667	2.2926	6.22	0.0000

Interpretação do modelo

O modelo ajustado finalmente é dado pela seguinte equação:

$$(\text{preço})_i = 2.32(\text{imposto})_i + 14.27(\text{areac})_i$$

Portanto, podemos dizer que com um aumento de 100 USD no imposto do imóvel, seu preço de venda aumenta em 2324.22 USD. Da mesma forma, com um aumento de 1000 pés quadrados de área construída, o preço do imóvel aumenta 14266.69 USD.

Análise de diagnóstico

Gráfico Quantil-Quantil

No gráfico quantil-quantil comparamos a distribuição dos resíduos com a distribuição normal. Esse gráfico indicará se a suposição de normalidade do modelo pode ou não ser aceita.

```
source("../funcoes/envel_norm.R")
envel_norm(modelo, imoveis) + xlab("Quantis da Normal") + ylab("Resíduo Studentizado")
```

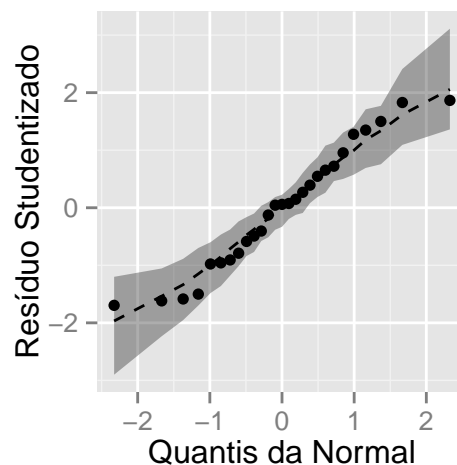


Figure 2: Gráfico Quantil-Quantil dos resíduos do modelo ajustado.

Como todos os pontos estão dentro da banda de confiança, podemos concluir que a suposição de normalidade é aceitável neste modelo.

Outros gráficos

Vamos analisar os demais gráficos de diagnóstico de regressão.

```
source("../funcoes/diag_norm.R")
diag_norm(modelo, imoveis)
```

Na figura 3 podemos ver que a observação 27 tem grande influência nas estimativas dos parâmetros. Os resíduos parecem ser aleatórios, indicando que as observações parecem ser independentes. Além disso, no

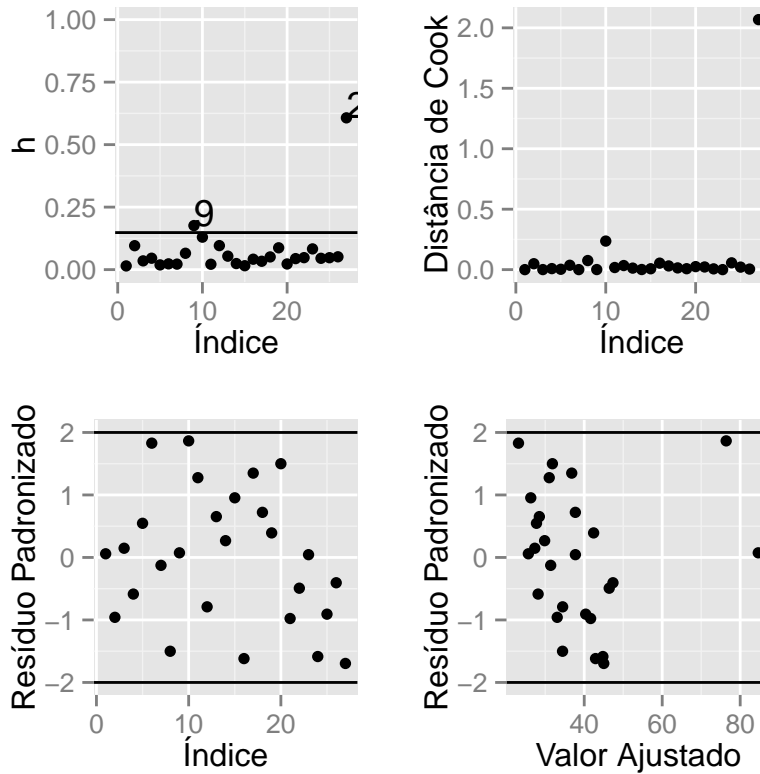


Figure 3: Gráfico Quantil-Quantil dos resíduos do modelo ajustado.

gráficos dos valores ajustados pelos resíduos, não parece existir nenhum padrão indicando que a amostra é homocedástica.

Vamos ajustar o modelo retirando a observação número 27, para verificar se existem grandes alterações nos parâmetros.

```
modelo.sem27 <- lm(preco ~ 0 + imposto + areac, data = imoveis[-27,])
```

	Est.	Est s/ obs 27
β_1	2.32	3.27
β_2	14.27	9.96

As estimativas dos parâmetros não se alteram fortemente, portanto optamos por não retirar esta observação da amostra.

Estimativa Intervalar para um imóvel fora da amostra

Supondo que um imóvel tenha um imposto representado por z_1 e uma área construída representada por z_2 , uma estimativa intervalar para o seu preço de venda (y) pode ser dada por:

$$\hat{y} = 2.32 * z_1 + 14.27 * z_2 \pm \sqrt{15.29(1 + z'(X^t X)^{-1} z)}$$

É importante que o vetor z esteja na região conjunta dos dados originais, usados para estimar os parâmetros do modelo. No caso a região é apresentada na figura abaixo:

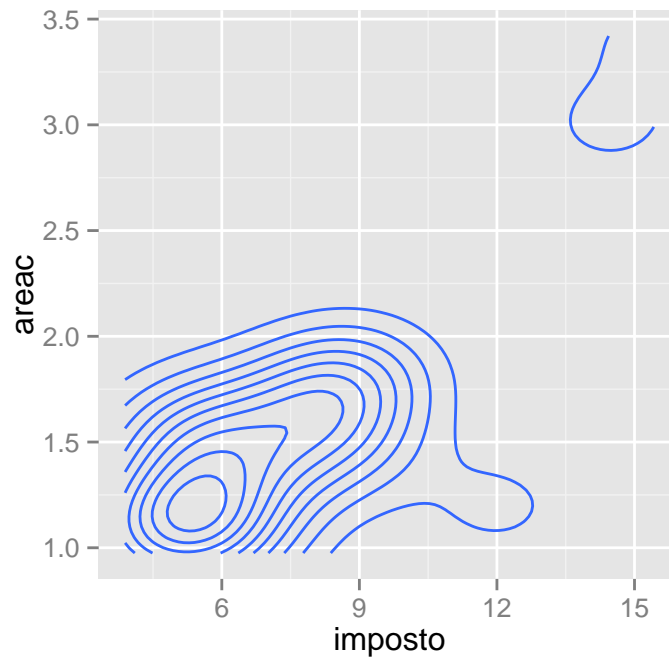


Figure 4: Região conjunta dos dados originais

Pontos fora da área azul seriam considerados extrapolações e é possível que as previsões feitas pelo modelo estejam erradas.