Exemplo

No arquivo **restaurante.dat** estão descritos os faturamentos anuais bem como os gastos com publicidade (em mil USD) de uma amostra aleatória de 30 restaurantes (Montgomery, Peck e Vining, 2001, p. 197-200). O objetivo principal é tentar relacionar o faturamento médio com o gasto com publicidade. Inicialmente faça uma análise descritiva dos dados, em particular o diagrama de dispersão entre as variáveis. Tente ajustar um modelo de regressão normal linear entre o faturamento e gastos e verifique através das técnicas de diagnóstico se existem afastamentos sérios das suposições feitas para o modelo.

Resolução

Inicialmente, faremos uma breve análise descritiva dos dados. Assim, apresentamos a Tabela 1 com as medidas-resumo referentes às variáveis **faturamento** e **gastos** e a Figura 1 com a densidade e o *boxplot* da variável **faturamento**.

Tabela 1: Medidas-resumo: faturamento médio e gastos com publicidade

	Min.	1º Quartil	Mediana	Média	3º Quartil	Máx.	d.p.
faturamento	72.34	117.70	147.00	147.10	180.30	218.70	42.13
gastos	3.00	8.93	12.46	12.13	15.19	19.50	5.12

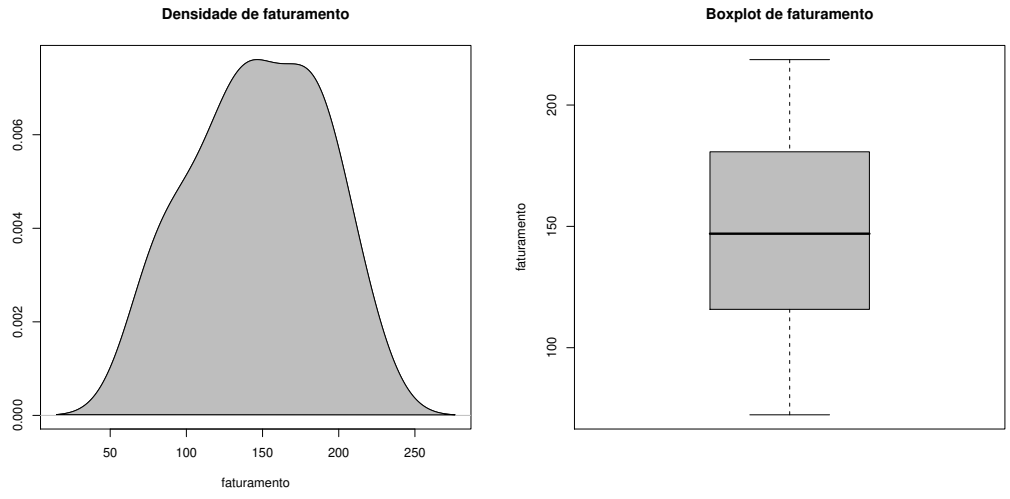

Figura 1: Densidade e boxplot da variável faturamento dos restaurantes

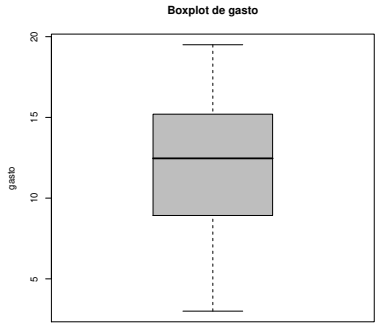

Figura 2: Boxplot da variável gastos com publicidade

Através da Tabela 1, observamos uma grande variabilidade do faturamento anual dos restaurantes, dado a grande amplitude do intervalo de variação do **faturamento** [72.37, 218.70]). No entanto, ao analisar a variável gastos com publicidade, notamos uma menor variabilidade, indicando que, aparentemente, não é necessário gastos muito maiores com publicidade para que haja grandes faturamentos. Ao observar a densidade e o boxplot da variável resposta **faturamento**, notamos que parece razoável ajustar um modelo simétrico, uma vez que as diferenças inter-quartis parecem muito próximas. Devido a ausência de *outliers* no *boxplot* da variável **faturamento**, faremos um ajuste normal linear. Note que o gráfico da densidade também nos sugere a ausência de caudas pesadas.

Na Figura 3, apresentamos o gráfico de dispersão entre as variáveis **faturamento** e **gastos**. Observe que, quanto maior o gasto com publicidade, maior é o faturamento anual do restaurante, indicando que existe uma relação linear entre as variáveis **faturamento** e **gastos** evindenciada pela grande correlaçao positiva entre as variáveis, calculada em 0.978. Além disso, veja que não são necessários grandes gastos com publicidade, para que haja grandes faturamentos.

Assim, ajustamos um modelo normal linear do tipo:

$$faturamento_i = \alpha + \beta\ gastos_i + \epsilon_i \tag{1}$$

em que $\epsilon_i \overset{ind}{\sim}$ Normal(0, σ^2) e $i = 1, ..., 30$ representam os restaurantes para os quais temos informações sobre o faturamento.

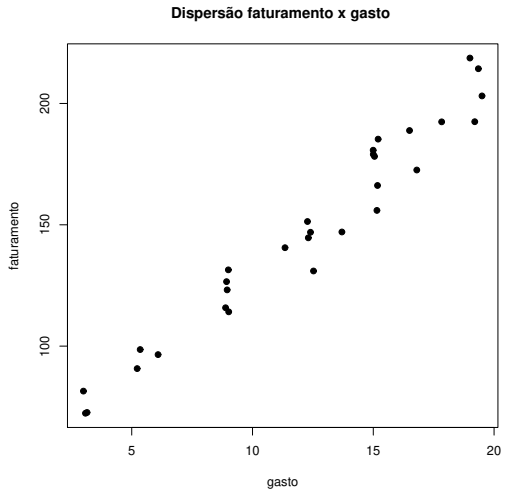

Figura 3: Gráficos de dispersão das variáveis faturamento e gastos

Os resultados do ajuste do modelo 1 estão descritos na Tabela 2. Podemos observar que a variável explicativa **gastos** é significativa ao nível de 1%, estando de acordo com a análise descritiva anterior em que havíamos notado que existe uma forte correlação entre a variável resposta **faturamento** e a variável explicativa **gastos**. A raiz do quadrado médio residual foi calculada em 8.999 em 28 graus de liberdade.

Tabela 2: Ajuste do modelo normal linear

	Estimativas	Erro Padrão	t-valor	p-valor
α	49.4434	4.2889	11.53	3.81e-12
β	8.0484	0.3265	24.65	$<$ 2e-16

Pelo gráfico de envelope da Figura 4, vemos que é razoável supor um modelo de regressão normal linear entre as variáveis **faturamento** e **gastos**, pois todos os pontos encontram-se dentro das bandas do envelope. O valor obtido para o coeficiente de determinação ao se supor normalidade foi $R^2 = 0.9544$, indicando que em torno de 95% da variação no faturamento está relacionada linearmente com os gastos em publicidade, sendo os outros 5% da variação resultantes de outros fatores não considerados (localização, tempo de existência do restaurante, etc.).

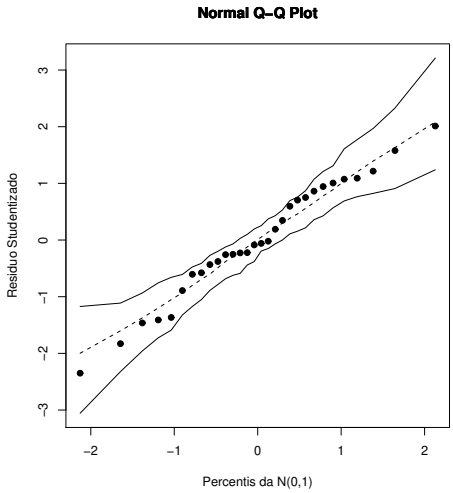

Figura 4: Envelope do ajuste normal

Na Figura 5, apresentamos diferentes gráficos que nos possibilitarão fazer uma análise de diagnóstico. Observe que o gráfico de alavanca destaca os pontos 1 (faturamento: 81.464, gasto:3), 2 (faturamento:72.661, gasto: 3.150) e 3 (faturamento:72.344, gasto: 3.085) como possíveis pontos de alavanca. Veja que estes pontos indicam restaurantes que têm baixo faturamento e pouco investem em publicidade. Além disso o ponto 1 tem um comportamento atípico, pois foi o restaurante que menos investiu em publicidade, mas teve um ganho superior aos restaurantes 2 e 3.

O restaurante 29 (faturamento: 218.715, gasto:19) aparece como um possível ponto influente e aberrante. Além disso, este ponto influi na suposição de homocedasticidade do modelo. Observe que este restaurante foi o que obteve maior faturamento anual, mas não foi o que mais investiu em publicidade.

O restaurante 15 (faturamento: 130.963, gasto: 12.525) aparece como um possível aberrante e também influi na suposição de homocedasticidade do modelo. Observe que este ponto também possui um comportamento atípico, pois investiu em publicidade 12.525 (mil USD), mas teve um faturamento de quem investiu em torno de 9 (mil USD).

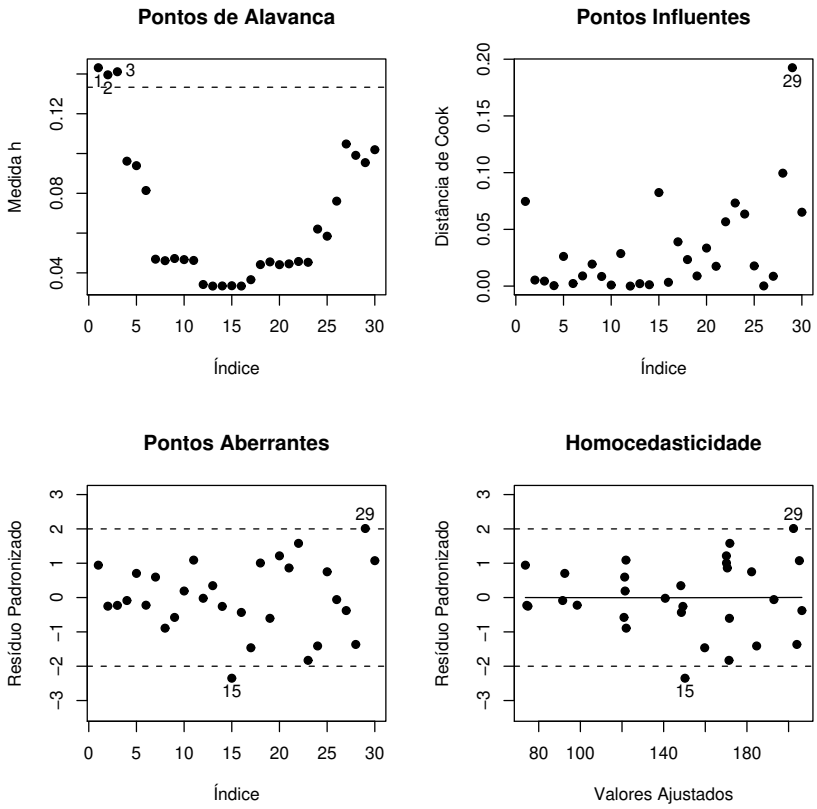

Figura 5: Diagnóstico do ajuste normal

Vejamos como se comportam as estimativas dos parâmetros do modelo normal linear, quando retiramos cada um dos pontos indicados anteriormente.

Analisando a Tabela 3, notamos que, de maneira geral, nenhum dos pontos retirados da análise influenciou significativamente as estimativas pontuais de α e β, tampouco mudaram as conclusões inferenciais do modelo completo.

Apesar de não mudar as conclusões inferenciais, notamos pelo gráfico dos resíduos padronizados versus valores ajustados na Figura 5 a presença de uma possível heteroscedasticidade no modelo. Neste momento, utilizaremos o método de Box-Cox para identificar uma possível transformação para ajudar a tornar o modelo homoscedástico.
FALTA CONTINUAR ESSA PARTE !!!!

Tabela 3: Ajuste do modelo normal após excluir observações críticas

		Estimativas	Erro Padrão	t-valor	p-valor
Dados Completos	α	49.4434	4.2889	11.53	3.81e-12
	β	8.0484	0.3265	24.65	$<$ 2e-16
sem 1	α	47.7962	4.6384	10.30	7.43e-11
	β	8.1590	0.3475	23.48	$<$ 2e-16
sem 2	α	49.882	4.699	10.62	3.87e-11
	β	8.019	0.352	22.78	$<$ 2e-16
sem 3	α	49.8428	4.7039	10.60	4.04e-11
	β	8.0217	0.3524	22.76	$<$ 2e-16
sem 15	α	49.983	3.986	12.54	9.02e-13
	β	8.059	0.303	26.59	$<$ 2e-16
sem 29	α	50.8238	4.1304	12.30	1.39e-12
	β	7.8850	0.3206	24.60	$<$ 2e-16
sem 15 e 29	α	51.2860	3.8112	13.46	3.18e-13
	β	7.9021	0.2955	26.74	$<$ 2e-16

Conclusões

Neste exercício procuramos estudar a associação entre o faturamento anual de um restaurante e os gastos anuais em publicidade. Vimos que, quanto maior o gasto em publicidade, maior é o faturamento anual do restaurante, existindo uma forte correlação entre estas duas variáveis, na medida em que, não é necessário grandes gastos em publicidade, para que haja grandes resultados no faturamento. Ao aplicarmos o modelo normal linear vimos que a análise de diagnóstico em ambos os modelos detectou como observações atípicas os restaurantes 1, 2, 3 e 15, 29. Estas observações representam casos extremos, em que ou houve um grande investimento em publicidade, mas não se obteve o retorno em faturamento esperado ou se obteve um grande lucro, sem se investir tanto em publicidade. Notamos que a retirada de cada um desses pontos em ambos os modelos não alterou as conclusões inferenciais, ainda que pontualmente as estimativas dos parâmetros mudassem um pouco quando comparados com as estimativas do parâmetro para a amostra completa. De maneira geral, o envelopes e o coeficiente de determinação foram razoáveis, indicando que o modelo normal linear teve um desempenho satisfatório na modelagem. No entanto parece haver indícios de heteroscedastidade, ... FALTA CONTINUAR ESSA PARTE !!!!