REGRESSÃO LINEAR

MOISÉS SALES

1. Página 108: Questão 20

No arquivo reg3.txt são descritas as seguintes variáveis referentes a 50 estados norte-americanos: (i) estado (nome do estado), (ii) pop (população estimada em julho de 1975), (iii) percap (renda percapita em 1974 em USD), (iv) analf (proporção de analfabetos em 1970), (v) expvida (expectativa de vida em anos 1969-70), (vi) crime (taxa de criminalidade por 100000 habitantes 1976), (vii) estud (porcentagem de estudantes que concluem o segundo grau 1970), (viii) ndias (número de dias do ano com temperatura abaixo de zero grau Celsius na cidade mais importante do estado) e (ix) area (área do estado em milhas quadradas). O objetivo do estudo é tentar explicar a expvida média usando um modelo de regressão normal linear dadas as variáveis explicativas percap, analf, crime, estud, ndias e dens, em que dens=pop/area. Inicialmente faça uma análise descritiva dos dados. Comente essa parte descritiva. Posteriormente, ajuste o modelo de regressão normal linear com todas as variáveis explicativas e através do método stepwise faça uma seleção de variáveis. Uma vez selecionado o modelo faça uma análise de diagnóstico e apresente as interpretações dos coeficientes estimados do modelo final.

Date: Fevereiro 2025.

1.1. **Resolução.** Será utilizado um nível de significância de 10% em todas as análises.

1.1.1. *Análise Descritiva.* Calculando alguma medidas descritivas para as variáveis, obtemos a tabela 1, na qual podemos perceber uma grande assimetria nas variáveis ”pop”e ”area”, que, obviamente, também está presente na variável ”dens”. Notamos uma grande curtose na variável densidade, o que causa caudas pesadas em sua distribuição, gerando, por sua vez, outliers; o mesmo pode ser observado na figura 2, que apresenta os boxplots das variáveis explicativas que serão utilizadas no modelo de regressão, em que a variável ”dens”apresenta 6 outliers; A variável que buscamos estudar, espvida, apresenta uma pequena assimetria negativa, apontando que os valores se concentram mais na cauda da direita, o que pode ser visto no histograma da figura 1, e uma baixa curtose, indicando a não existência de outliers, com o auxílio do boxplot na figura 1, confirmamos a não existência de outliers.

	Media	Mediana	D.P.	Intervalo	Assimetria	Curtose
pop	4246.40	2838.50	4464.50	20833.00	1.90	3.80
percap	4435.80	4519.00	614.50	3217.00	0.20	0.20
analf	1.20	0.90	0.60	2.30	0.80	-0.50
expvida	70.90	70.70	1.30	5.60	-0.20	-0.70
crime	7.40	6.80	3.70	13.70	0.10	-1.20
estud	53.10	53.20	8.10	29.50	-0.30	-0.90
ndias	104.50	114.50	52.00	188.00	-0.40	-0.90
area	72862.20	54277.00	89719.80	565383.00	3.60	16.20
dens	0.20	0.10	0.40	2.70	4.60	24.40

TABELA 1. Medidas descritivas das variáveis

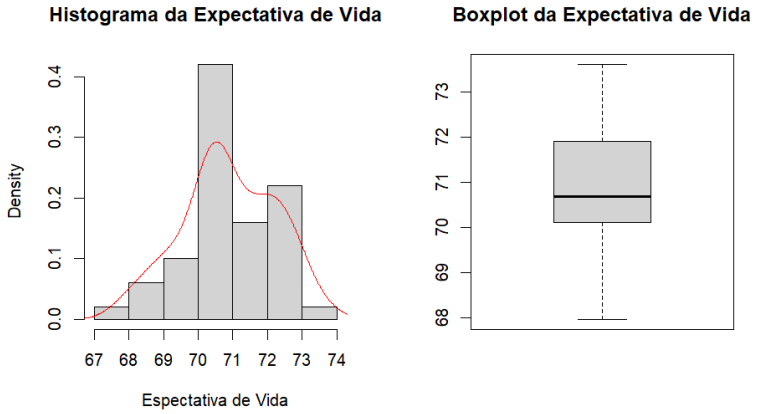


FIGURA 1. Histograma e BoxPlot da Expectativa de Vida

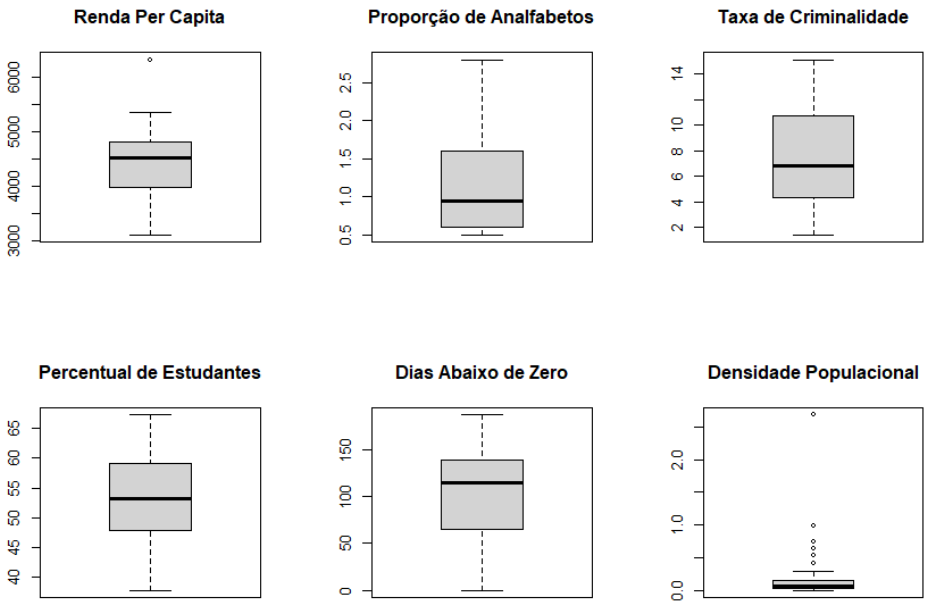


FIGURA 2. Boxplot das variáveis explicativas

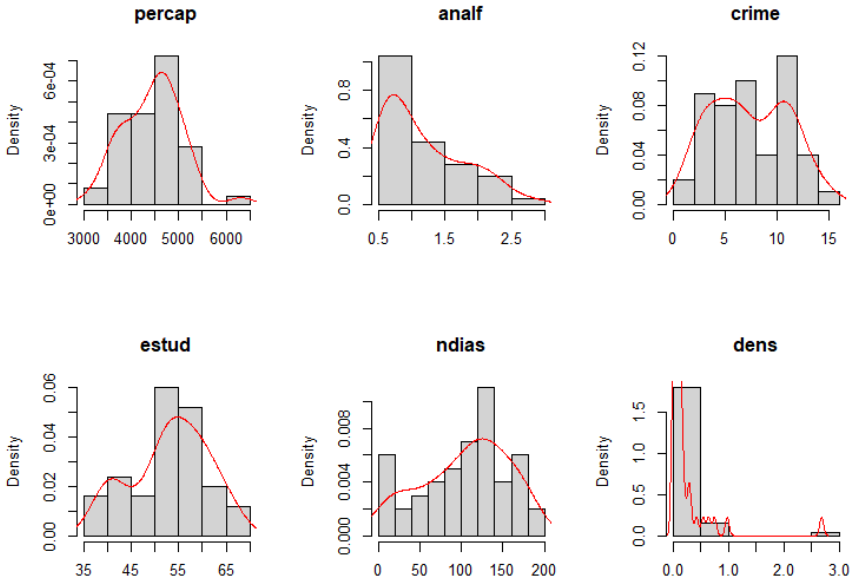


FIGURA 3. Histograma das variáveis explicativas

Utilizando o teste de Tukey para encontrar outliers, a variável ”percap”apresentou 1 outlier: Alaska, e a variável ”dens”apresentou 6 outliers: Carolina do Sul, Nova Jersey, Massachusetts, Connecticut, Virginia e Maryland.

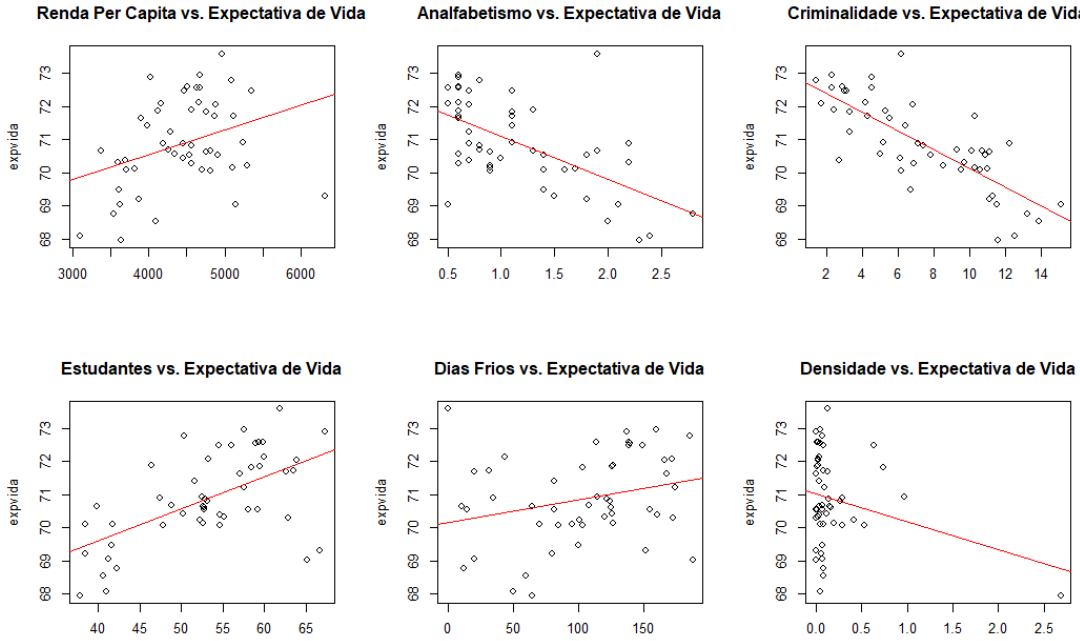


FIGURA 4. Gráficos de Dispersão das variáveis explicativas

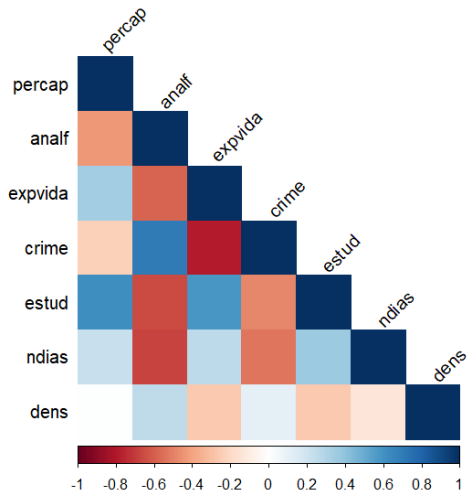


FIGURA 5. Matriz de Correlação das variáveis explicativas

Analisando o gráfico 4, podemos notar que existe uma correlação negativa entre analfabetismo e expectativa de vida, ou seja, quanto maior o nível de analfabetismo em um estado, menor sua expectativa de vida. O mesmo pode ser dito para o gráfico de criminalidade vs expectativa de vida, o que é esperado. Esses resultados podem ser verificados com o auxílio da figura5, que aponta a correlação entre todas as variáveis explicativas. Podemos notar, através dessa matriz, que existe uma correlação negativa entre o número de dias frios e o nível de analfabetismo, e como o analfabetismo possui uma correlação negativa com a renda per capita, pessoas que moram em cidades mais frias tendem a ter uma renda per capita maior.

1.1.2. *Modelo de Regressão.* Calculando o modelo com todas as variáveis explicativas, temos:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.3534	1.3755	51.87	0.0000
percap	0.0002	0.0002	0.96	0.3409
analf	-0.0528	0.3297	-0.16	0.8735
crime	-0.2876	0.0414	-6.95	0.0000
estud	0.0304	0.0207	1.47	0.1495
ndias	-0.0075	0.0028	-2.71	0.0097
dens	-0.5118	0.2764	-1.85	0.0709

TABELA 2. Modelo com todas as variáveis

Utilizando o método stepwise(forward e backward) para a seleção de variáveis, o modelo resultante é dado por:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.5015	0.9984	71.61	0.0000
crime	-0.2859	0.0360	-7.95	0.0000
estud	0.0436	0.0153	2.85	0.0065
ndias	-0.0071	0.0024	-2.98	0.0047
dens	-0.4568	0.2620	-1.74	0.0880

TABELA 3. Modelo com as variáveis selecionadas

1.1.3. *Análise de Diagnóstico.* Realizando o teste de Shapiro-Wilk para testar a normalidade dos resíduos, temos que o p-valor = 0.2284, ou seja, não rejeitamos a hipótese de normalidade dos resíduos. O gráfico do envelope da normal nos resíduos, figura 6, apresenta o mesmo resultado.

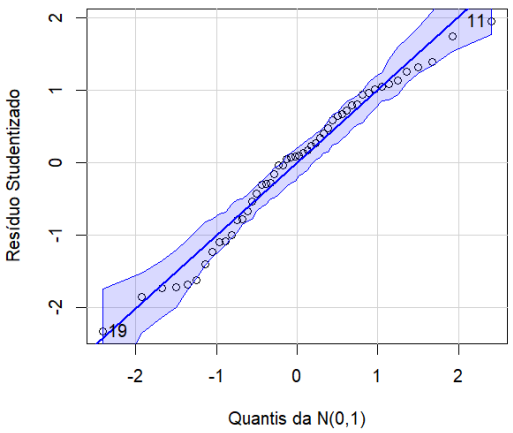


FIGURA 6. Gráfico Quantil-Quantil para os Resíduos Studentizados

A figura 7, nos ajuda a verificar a suposição de homocedasticidade, apenas um ponto se distancia da banda de confiancia, que é o estado de Maryland, um outlier, porém, ao realizarmos o teste de Breusch-Pagan, para a heterocedasticidade, resulta em um p-valor = 0.5254, ou seja, não rejeitamos a hipótese de homocedasticidade.

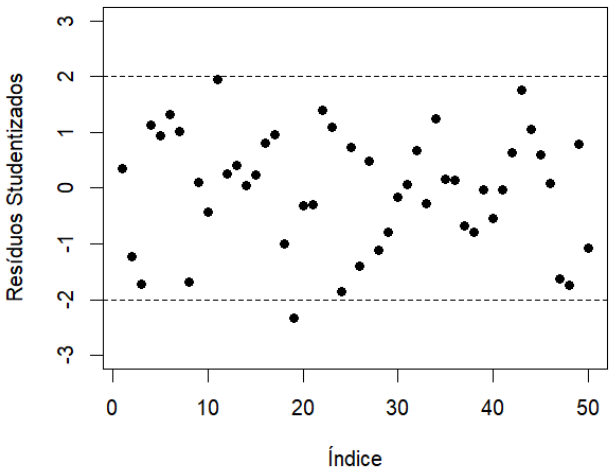
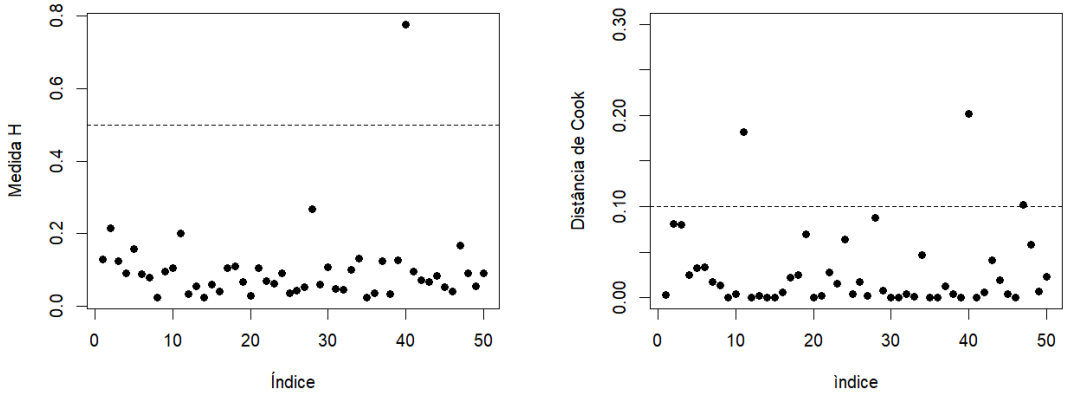


FIGURA 7. Resíduos vs índices

Pela figura 8a, o estado de índice 40, Carolina do Sul, que é um outlier na variável ”dens”, é um ponto de alavanca, ou seja, é um ponto que possui um alto valor nas variáveis explicativas. Esse ponto pode ser um ponto influente e afetar nas estimativas dos parâmetros. O que é confirmado com o auxílio da figura 8b, que apresenta a distância de Cook dos resíduos, temos que a Carolina do Sul, ponto com índice 40, e o Hawaii, índice 12, são pontos influentes.



(A) Gráfico de Alavanca (B) Distância de Cook

FIGURA 8. Gráficos de Alavanca e Influência referentes a modelo normal.

1.1.4. *Interpretação dos coeficientes estimados.* Os coeficientes estimados do modelo final são dados pela seguinte tabela: Em que o Intercepto = 71.5 representa a média de

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.5015	0.9984	71.61	0.0000
crime	-0.2859	0.0360	-7.95	0.0000
estud	0.0436	0.0153	2.85	0.0065
ndias	-0.0071	0.0024	-2.98	0.0047
dens	-0.4568	0.2620	-1.74	0.0880

anos de Expectativa de Vida, a estimativa do coeficiente da variável crime = -0.2859 representa o quanto de influência o aumento de uma unidade na variável crime influencia, em média, na expectativa de vida, ou seja, um aumento de unidade da variável crime afeta, em média, a expectativa de vida em -0.2859 anos. O mesmo pensamento se aplica a variável estudo, um aumento de unidade da variável estudo aumenta, em média, 0.0436 anos na Expectativa de Vida. O mesmo raciocínio se aplica a todas os outros coeficientes estimados.