UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

ANÁLISE DA RELAÇÃO ENTRE A RENDA MENSAL E VARIÁVEIS SOCIOECONÔMICAS, DEMOGRÁFICAS E CLIMÁTICAS EM VÁRIOS PAÍSES: UM ESTUDO DE REGRESSÃO LINEAR

Salvador
2023

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

ANÁLISE DA RELAÇÃO ENTRE A RENDA MENSAL E VARIÁVEIS SOCIOECONÔMICAS, DEMOGRÁFICAS E CLIMÁTICAS EM VÁRIOS PAÍSES: UM ESTUDO DE REGRESSÃO LINEAR

Atividade de laboratório apresentada ao Instituto de Matemática e Estatística da Universidade Federal da Bahia como parte das exigências da disciplina Análise de Regressão ministrada pela professora Dra. Edleide de Brito.

Salvador
2023

SUMÁRIO

1 Introdução

A análise da relação entre a renda mensal e outras variáveis é essencial para entender os fatores que influenciam o bem-estar das pessoas em diferentes países. O objetivo deste projeto é investigar a relação entre a renda mensal e um conjunto de variáveis socioeconômicas, demográficas e climáticas em vários países por meio de uma análise de regressão linear simples. As variáveis preditoras incluem expectativa de vida masculina e feminina, taxa de natalidade, taxa de mortalidade, quociente de inteligência, gastos com educação por habitante, temperatura máxima diária, índice de custo e poder de compra indexado por país e continente.

Inicialmente, será realizada uma análise descritiva dos dados para observar as medidas de tendência central e variabilidade das variáveis, bem como a presença de valores atípicos. Serão examinadas as relações entre as variáveis, principalmente a relação de cada variável com a renda mensal. Se houver linearidade, isso indicará a possibilidade de um modelo de regressão linear explicar ou prever a renda mensal.

Será realizado um modelo de regressão linear simples para cada variável em relação à renda mensal, a fim de verificar se o modelo é significativo e obter interpretações para cada modelo. Em seguida, será efetuada uma análise dos resíduos em pelo menos um dos modelos, para verificar se a hipótese de normalidade, homocedasticidade e independência dos erros está sendo satisfeita pelo modelo.

Para isso, serão utilizados gráficos da densidade empírica dos resíduos, *qq-plot* e o teste de normalidade de Shapiro-Wilk para avaliar a normalidade dos resíduos; gráficos de dispersão dos resíduos em relação aos valores ajustados e o teste de Breush-Pagan para verificar a homocedasticidade; bem como o teste de Durbin-Watson para verificar a independência dos erros.

Todas as análises necessárias serão realizadas com o auxílio do software R, (R Core Team, 2022). Com base nos resultados obtidos, serão tiradas conclusões sobre a capacidade do modelo em explicar a relação entre as variáveis socioeconômicas, demográficas e climáticas e a renda mensal.

2 Materiais e métodos

2.1 Descrição dos dados

A base de dados contém informações socioeconômicas, demográficas e climáticas de 82 países, disponíveis no site `https://www.dadosmundiais.com/`. Os dados foram obtidos por meio da combinação de bases de dados contendo informações sobre a expectativa de vida, quociente de inteligência e custo de vida de cada país.

As variáveis presentes no conjunto de dados são:

Variável	**Descrição**
País	Nome do país
QI	Quociente de inteligência
Despesas com educação	Gastos do estado com educação em dólares
Temperatura máxima diária	Temperatura máxima diária em graus celsius
Renda mensal	Renda mensal, em dolar, calculada a partir da renda nacional bruta por habitante
Expectativa de vida masculina	Expectativa de vida masculina em anos
Expectativa de vida feminina	Expectativa de vida feminina em anos
Taxa de natalidade	Taxa de nascimento por 1000 habitantes
Taxa de mortalidade	Taxa de morte por 1000 habitantes
Continente	Continente em que os países estão localizados

Tabela 2.1: Variáveis disponíveis no conjunto de dados

3 Resultados

3.1 Análise Descritiva

Como pode ser visto na Tabela 3.1, os dados não parecem apresentar simetria se avaliarmos a distância entre os valores de média, moda e mediana. As variáveis da renda mensal e das despesas com educação têm uma alta variabilidade, o que pode ser resultado da diferença de renda e dos gastos com a educação entre os países.

Tabela 3.1: Sumarização das variáveis socioeconômicas, demográficas e climáticas dos países em 2019

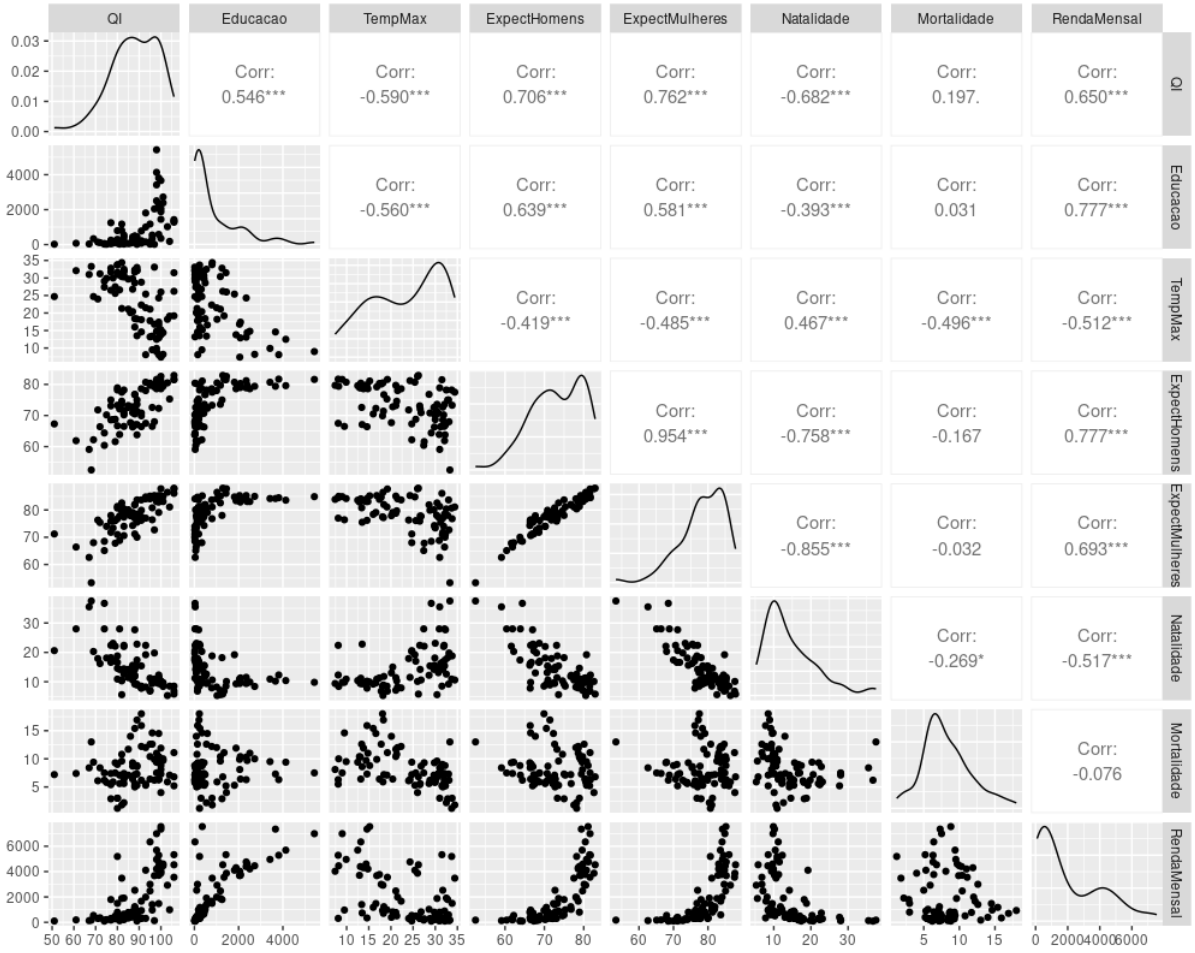
	QI	Desp. Edu.	Temp.	Renda Men.	Exp.M.	Exp.F.	Nat.	Mort.
Mínimo	51,0	14,0	7,4	92,0	52,5	53,3	5,3	1,2
1º quartil	81,0	116,8	16,7	337,8	68,6	75,6	9,6	6,1
Mediana	88,5	322,0	25,1	950,5	73,6	79,1	12,0	7,4
Moda	99,0	76,0	31,7	173,0	78,6	85,3	22,4	6,4
Média	88,1	857,8	23,3	1988,6	73,2	78,5	14,5	8,2
3º quartil	98,0	1257,5	31,0	3700,2	79,2	83,7	18,1	10,0
Máximo	106,0	5425,0	34,4	7550,0	82,9	88,0	37,5	18,0
Variância	121,4	1275572,0	65,2	4191637,0	43,6	41,7	50,3	11,7

Observando a Figura 3.1, as variáveis no geral apresentam correlações significativas entre si. Quando as variáveis independentes tem um coeficiente de correlação de *Pearson* com valor absoluto alto com a variável dependente, isto é algo positivo para o modelo de regressão linear. Entretanto, quando as variáveis dependentes apresentam uma correlação absoluta alta entre si, pode haver problema de multicolinearidade no modelo de regressão.

As variáveis QI, expectativa de vida masculina, expectativa de vida feminina e natalidade aparentam ter uma relação não linear, mas exponencial com a variável resposta, renda familiar. Pensando em um modelo de regressão linear, isto pode claramente gerar problemas nos resíduos, como a violação da suposição de homoscedasticidade — provavelmente, com a variância crescendo para maiores valores ajustados da variável resposta.

A única variável que aparenta ter uma relação linear notável com a variável renda

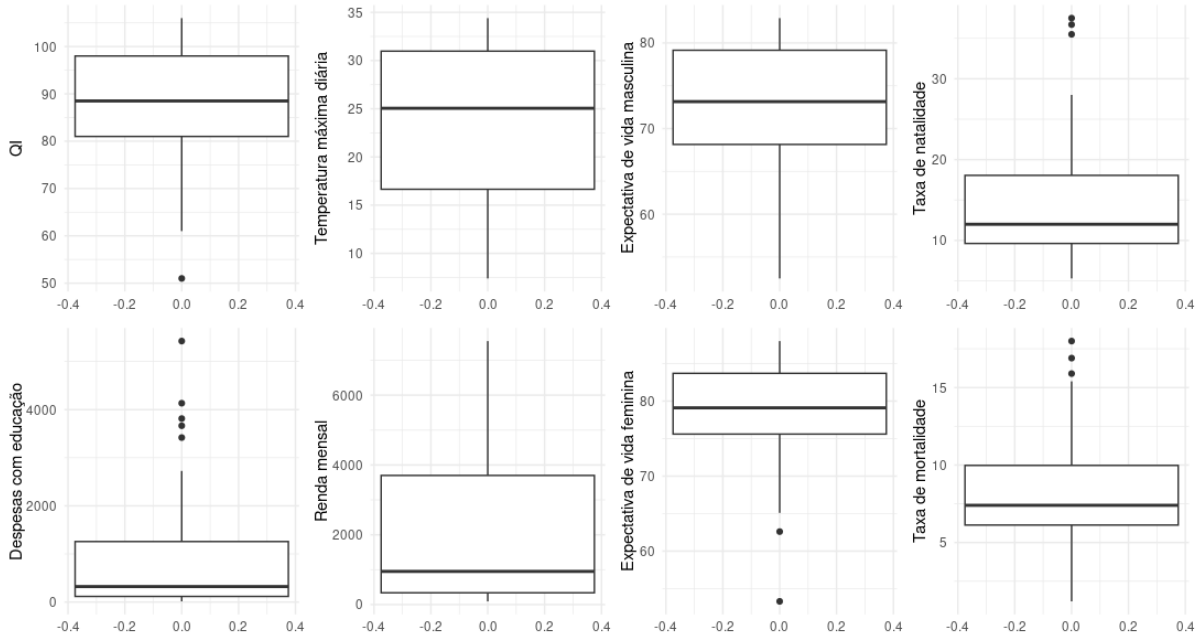
Figura 3.1: Gráfico de dispersão, densidade e correlação das variáveis socioeconômicas, demográficas e climáticas dos países em 2019



familiar é as despesas com educação, apesar de ter pontos que fogem dessa linearidade. A variável temperatura máxima diária tem uma fraquíssima relação linear negativa com a renda familiar. A variável taxa de mortalidade não parece ter nenhum tipo de relação com a renda familiar.

Como citado anteriormente, há uma grande variabilidade na renda mensal. Observando o gráfico de *boxplots* da Figura 3.2, é possível notar que essa grande variabilidade está concentrada acima da mediana. O *boxplot* do QI, da taxa de natalidade, das despesas com educação, da expectativa de vida feminina e da taxa de mortalidade indicou presença de *outliers*, isso pode gerar problemas no que tange ao modelo de regressão linear simples. Entretanto, retirar estes valores da análise pode ser problemático, pois cada observação representa um país e certamente o fato destes valores serem extremos nos traz alguma informação sobre um problema social: a desigualdade presente entre os diferentes países do mundo.

Figura 3.2: *Boxplots* das variáveis socioeconômicas, demográficas e climáticas dos países em 2019



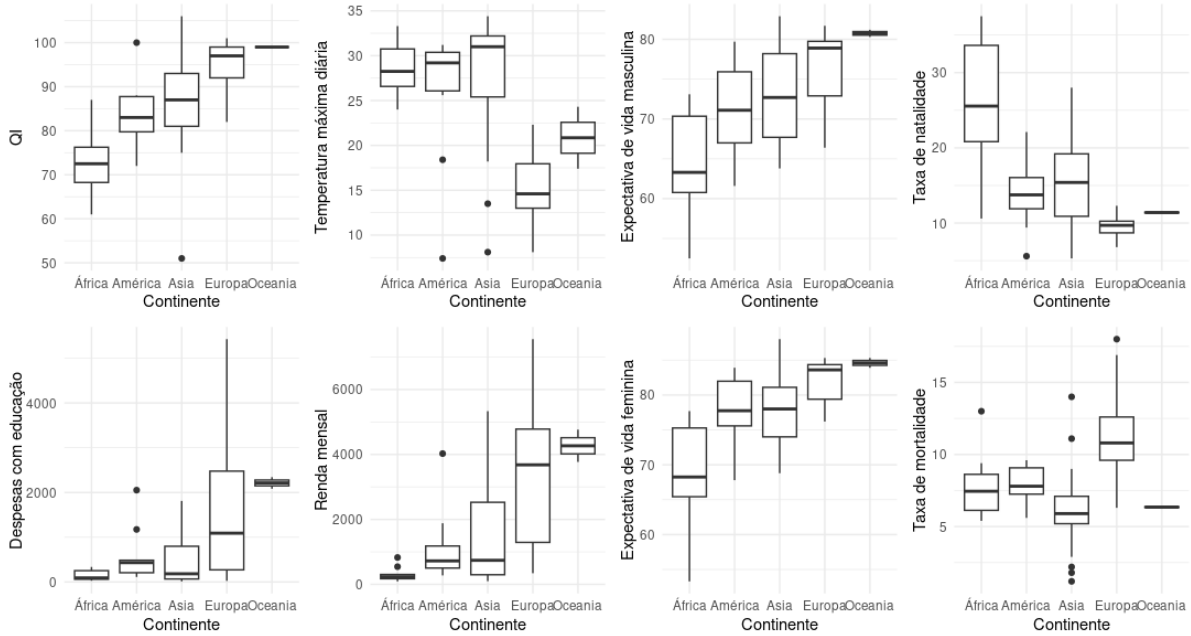
A Oceania tem apenas dois países: Austrália e Nova Zelândia, portanto é mais complicado analisar as características de dispersão das variáveis em relação a esse continente. Mas, o é possível observar no gráfico 3.3, é que em relação ao QI, a Oceania apresenta a maior mediana, seguida por Europa, enquanto a África apresenta o menor QI mediano. A Europa apresenta maior variabilidade do QI entre o 1º quartil e a mediana, o que indica uma assimetria a esquerda.

Em relação a temperatura máxima diária, a Ásia apresenta a maior mediana, com bastante variabilidade entre o 1º quartil e a mediana, indicando que há uma discrepância entre a temperatura dos países da Ásia. Há presença de valores atípicos na América e na Ásia, o que é esperado devido a extensão longitudinal desses continentes. A Europa é o continente com as menores temperaturas máximas.

A expectativa de vida masculina e feminina apresenta maiores valores na Oceania e na Europa. A África apresenta os menores valores de expectativa de vida, com uma alta variabilidade entre o 1º quartil e a mediana. Em todos os continentes a mediana da expectativa de vida feminina é maior que a mediana da expectativa de vida masculina.

As despesas com educação apresentam uma distribuição entre os continentes parecida com a renda mensal, tendo uma alta variabilidade na Europa e com uma baixa variabilidade e menores valores na África e na América. Um ponto interessante é que a América apresenta valores de mediana das despesas com educação e da renda mensal maiores do que a mediana das da Ásia. Mas, a Ásia apresenta uma maior variabilidade e maiores

Figura 3.3: *Boxplots*, de acordo com o continente, das variáveis socioeconômicas, demográficas e climáticas dos países em 2019



valores acima da mediana do que a América.

A mediana da taxa de natalidade na África é consideravelmente maior do que em outros continentes, assim como a sua variabilidade. A Europa apresenta uma taxa de natalidade com mediana bem pequena e com pouca variabilidade. Entretanto, não é possível estabelecer relação entre as variáveis taxa de natalidade e taxa de mortalidade, pois a Europa apresenta uma maior mediana da taxa de mortalidade enquanto a Ásia apresenta uma menor mediana da taxa de mortalidade. Essa variável apresenta diversos valores atípicos, principalmente na Ásia.

3.2 Modelos de regressão linear

A Tabela 3.2, apresenta a sumarização das regressões lineares em que as variáveis de expectativa de vida, quociente de inteligência e custo de vida são utilizadas para predizer a renda mensal.

Tabela 3.2: Sumarização das regressões lineares em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal dos países em 2019

Coeficientes	Estimativa	EP	Est. T	Valor-p	R^2
Intercepto	-8653,50	1400,47	-6,18	<0,001	0,42
QI	120,85	15,78	7,66	<0,001	
Intercepto	811,53	183,43	4,42	<0,001	0,60
Desp. Edu.	1,42	0,13	10,91	<0,001	
Intercepto	5018,06	600,42	8,36	<0,001	0,26
Temperatura	-129,88	24,34	-5,34	<0,001	
Intercepto	-15658,05	1602,44	-9,77	<0,001	0,60
Exp. masc.	241,14	21,81	11,06	<0,001	
Intercepto	-15259,26	2015,06	-7,57	<0,001	0,48
Exp. fem.	219,72	25,59	8,59	<0,001	
Intercepto	4158,21	446,64	9,31	<0,001	0,27
Natalidade	-149,19	27,64	-5,40	<0,001	
Intercepto	2361.82	594,71	3,971	<0,001	0.006
Mortalidade	-45,34	66,79	-0,68	0,50	

A interpretação de cada um dos modelo é:

- Modelo de regressão com QI como variável independente: este modelo indica que para cada unidade de aumento no quociente de inteligência, a renda mensal aumenta em média 120,85 dólares. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 42% da variação na renda mensal pode ser explicada pela variação no quociente de inteligência;

- Modelo de regressão com despesas com educação como variável independente: este modelo indica que para cada unidade de aumento nas despesas com educação, a renda mensal aumenta em média 1,42 dólares. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 60% da variação na renda mensal pode ser explicada pela variação nas despesas com educação;

- Modelo de regressão com temperatura máxima diária como variável independente: este modelo indica que para cada unidade de aumento na temperatura máxima

diária, a renda mensal diminui em média 129,88 dólares. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 26% da variação na renda mensal pode ser explicada pela variação na temperatura máxima diária;

- Modelo de regressão com expectativa de vida masculina como variável independente: este modelo indica que para cada unidade de aumento na expectativa de vida masculina, a renda mensal aumenta em média 241,14 dólares. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 60% da variação na renda mensal pode ser explicada pela variação na expectativa de vida masculina;

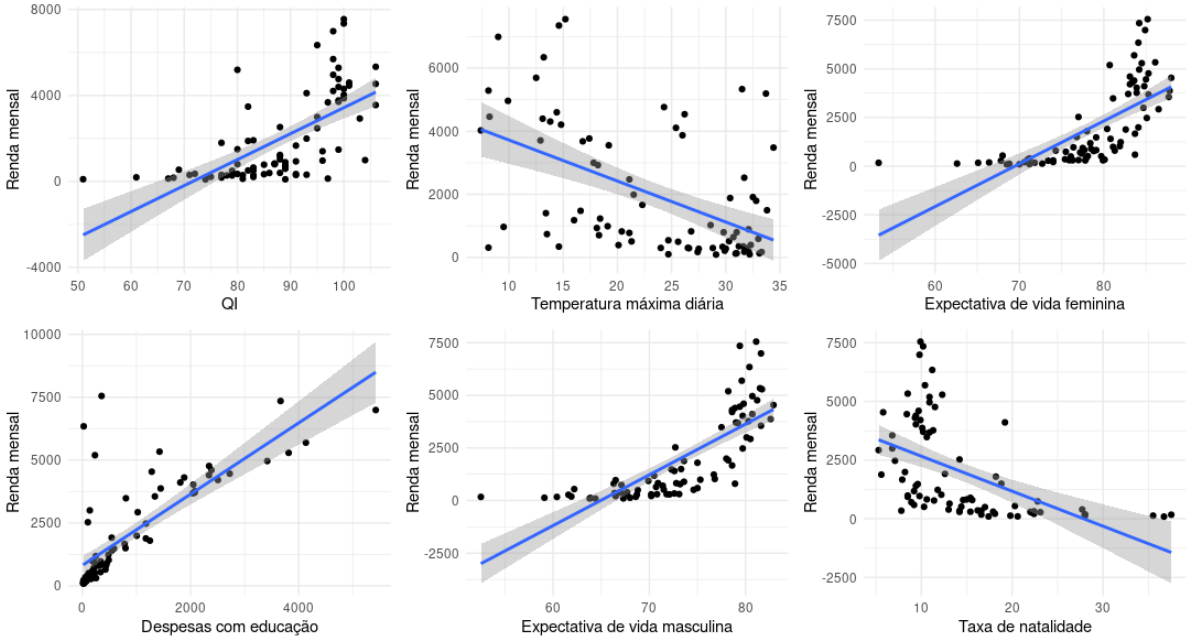
- Modelo de regressão com expectativa de vida feminina como variável independente: este modelo indica que para cada unidade de aumento na expectativa de vida feminina, a renda mensal aumenta em média 219,72 dólares. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 48% da variação na renda mensal pode ser explicada pela variação na expectativa de vida feminina;

- Modelo de regressão com taxa de natalidade como variável independente: este modelo indica que para cada unidade de aumento na taxa de natalidade, a renda mensal diminui em média 149,19 dólares. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 27% da variação na renda mensal pode ser explicada pela variação na taxa de natalidade.

- Modelo de regressão com taxa de mortalidade como variável independente: este modelo indica que para cada unidade de aumento na taxa de natalidade, a renda mensal diminui em média 45,34 dólares. O valor-p sugere que este coeficiente não é significativo e o R^2 indica que cerca de 0,6% da variação na renda mensal pode ser explicada pela variação na taxa de mortalidade. Portanto, a taxa de mortalidade não é uma variável que explica linearmente bem a renda mensal.

A Figura 3.4 apresenta a reta estimada em cada modelo de regressão apresentado. É possível notar que a reta estimada com as variáveis que denotam QI, expectativa de vida feminina, expectativa de vida masculina e taxa de natalidade não parecem acompanhar os pontos observados. Pois, há valores muito altos e muito baixos na renda mensal, afetando negativamente a precisão do modelo, aparentando ter uma relação exponencial com a variável resposta

Figura 3.4: Gráfico de dispersão e curva de regressão linear ajustada aos dados de países em 2019



Devido a essa relação exponencial com a renda mensal, nos modelos com a variável explicativa sendo QI, temperatura máxima diária ,taxa de natalidade, expectativa de vida feminina e masculina, a transformação logarítmica da renda mensal pode ser considerada. Pois, ajudará a reduzir a escala dos valores e tornar a relação entre as variáveis mais linear.

3.3 Modelos de regressão linear transformado

A Tabela 3.3, apresenta uma sumarização das regressões lineares em que as variáveis de quociente de inteligência, temperatura, expectativa de vida e natalidade são utilizadas para predizer o logaritmo da renda mensal.

Como o modelo com as despesas com educação pareceu bem predito e sua relação com a variável resposta mais linear, não foi considerado a transformação da variável dependente para esse modelo. Além disso, não foi considerada a transformação do modelo com a taxa de mortalidade, pois o modelo não foi significativo e pouco explicou a variação observada na renda mensal.

Tabela 3.3: Sumarização das regressões lineares em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal dos países em 2019

Coeficientes	Estimativa	Exp(Est.)	EP	Est. T	Valor-p	R^2
Intercepto	-0,48	0,62	0,79	-0,61	0,54	0.53
QI	0,08	1,09	0,01	9,45	<0,001	
Intercepto	8,86	7061,20	0,37	23,93	<0,001	0,28
Temperatura	-0,08	0,92	0,02	-5,54	<0,001	
Intercepto	-5,34	0,01	0,79	-6,76	<0,001	0,75
Exp. masc.	0,17	1,18	0,01	15,58	<0,001	
Intercepto	-5,96	0,003	0,97	-6,15	<0,001	0,69
Exp. fem.	0,16	1,18	0,01	13,33	<0,001	
Intercepto	8,78	6500,73	0,23	38,34	<0,001	0,50
Natalidade	-0,13	0,88	0,01	-9,01	<0,001	

A interpretação de cada um dos modelo é:

- Modelo de regressão com QI como variável independente: este modelo indica que para cada unidade de aumento no quociente de inteligência, a renda mensal aumenta em média 9%. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 53% da variação na renda mensal pode ser explicada pela variação no quociente de inteligência;

- Modelo de regressão com temperatura máxima diária como variável independente: este modelo indica que para cada unidade de aumento na temperatura máxima diária, a renda mensal diminui em média 8%. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 28% da variação na renda mensal pode ser explicada pela variação na temperatura máxima diária;

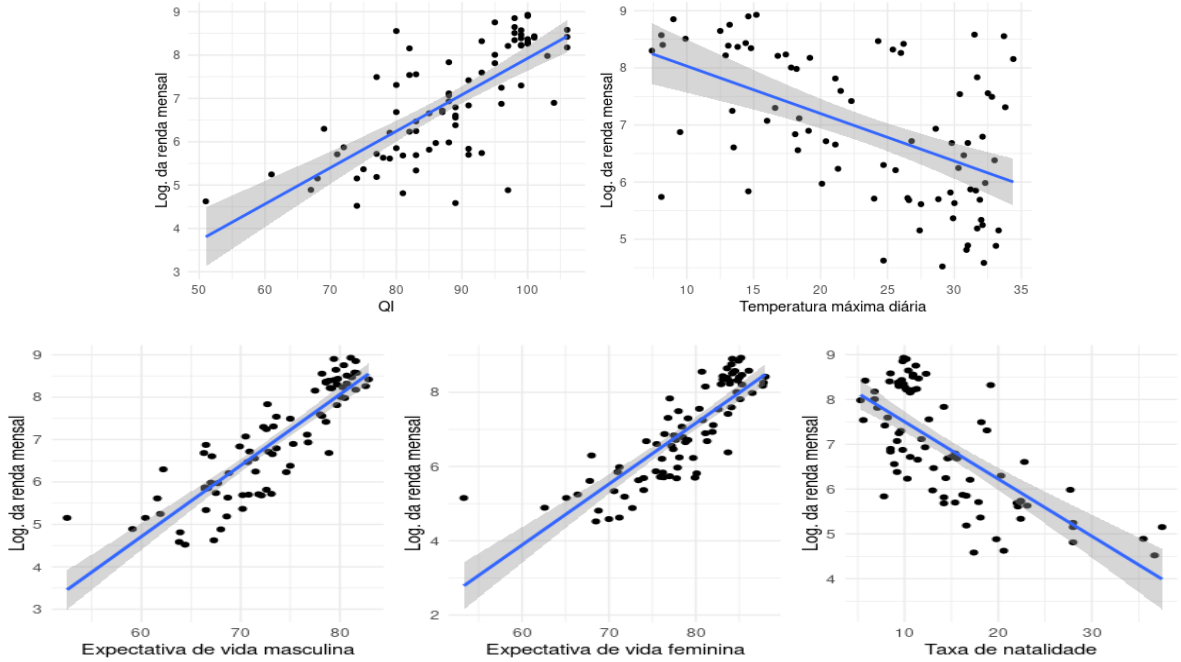
- Modelo de regressão com expectativa de vida masculina como variável independente: este modelo indica que para cada unidade de aumento na expectativa de vida masculina, a renda mensal aumenta em média 18%. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 75% da variação na renda mensal pode ser explicada pela variação na expectativa de vida masculina;

- Modelo de regressão com expectativa de vida feminina como variável independente: este modelo indica que para cada unidade de aumento na expectativa de vida feminina, a renda mensal aumenta em média 18%. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 69% da variação na renda mensal pode ser explicada pela variação na expectativa de vida feminina;

- Modelo de regressão com taxa de natalidade como variável independente: este modelo indica que para cada unidade de aumento na taxa de natalidade, a renda mensal diminui em média 12% . O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 50% da variação na renda mensal pode ser explicada pela variação na taxa de natalidade.

A Figura 3.5 ilustra as retas estimadas em cada modelo de regressão, após a aplicação da transformação logarítmica na variável dependente.

Figura 3.5: Gráfico de dispersão e curva de regressão linear ajustada aos dados de países em 2019



É possível notar que o QI, a expectativa de vida masculina, a expectativa de vida feminina e a taxa de natalidade parecem ser bons preditores da renda mensal nos países

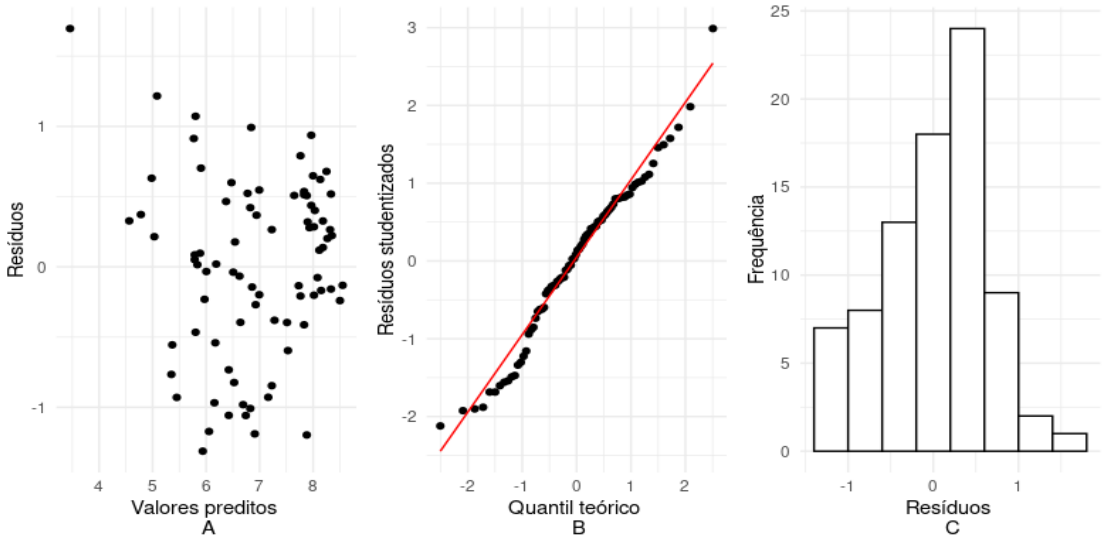
considerados. Pois, a reta estimada consegue acompanhar a relação observada nos pontos. Ao contrário da temperatura máxima diária, que não apresenta relação linear tão clara e grande parte dos pontos fogem do intervalo de confiança.

3.4 Análise dos resíduos

A Figura 3.6 apresenta a análise dos resíduos do modelo em que a expectativa de vida masculina é utilizada para predizer a renda mensal.

Em relação a Figura 3.6 (A), é possível notar que a dispersão dos resíduos permanece uniforme em relação aos valores preditos. Embora o teste de *Breusch-Pagan* tenha apontado a rejeição da hipótese de homocedasticidade (valor-p = 0,002), a avaliação gráfica sugere que a variância dos resíduos pode ser constante. Isso se deve ao fato de que o resultado do teste pode ter sido influenciado por uma única observação discrepante, e não necessariamente reflete uma falta de homogeneidade na variância dos resíduos.

Figura 3.6: Gráfico dos resíduos versus valores ajustados, resíduos studentizados versus quantil teórico e histograma dos resíduos da regressão linear em que a expectativa de vida masculina é utilizada para predizer a renda mensal de países em 2019



A Figura 3.6 (B), que apresenta o *qq-plot* dos resíduos studentizados, mostra que a suposição de normalidade não é violada pelos resíduos, embora um valor discrepante pareça afetar esse pressuposto. Esse valor discrepante, torna a cauda da distribuição desses resíduos mais pesada à direita, conforme ilustrado na Figura 3.6 (C). O teste de *Shapiro-Wilk* apresentou um valor-p de 0,19, não permitindo a rejeição da hipótese de normalidade a um nível de significância de 5%, corroborando com a análise gráfica realizada.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente 0,06, apon-

tando que não há autocorrelação serial nos resíduos a um nível de 5%. Ou seja, não há a violação do pressuposto de independência.

Embora seja importante levar em consideração os resultados dos testes estatísticos, a avaliação gráfica dos resíduos é mais confiável do que os testes. Pois, os testes podem não ser sensíveis o suficiente para detectar violações dos pressupostos em conjuntos de dados pequenos ou com pouca variação, ou podem ser hipersensíveis para rejeitar a hipótese de não violação do pressuposto se há presença de valores mais extremos, como foi o caso nessa análise de resíduos realizada. Com base nos resultados dos testes estatísticos fornecidos e sobretudo da avaliação gráfica, pode-se concluir que os resíduos atendem a todos os pressupostos.

4 Conclusão

A análise descritiva realizada permitiu uma visão geral sobre as características das variáveis estudadas. Observou-se que a Oceania e a Europa apresentam melhores indicadores de expectativa de vida, renda mensal e quociente de inteligência (QI), enquanto a África apresenta os menores valores para essas mesmas variáveis.

As variáveis QI, despesas com educação, expectativa de vida masculina, expectativa de vida feminina e taxa de natalidade apresentaram alguma relação com a renda mensal, sendo que apenas despesas com educação aparentava ter uma relação linear com a renda mensal. Diferentemente das variáveis que denotam temperatura máxima diária e taxa de mortalidade, que apresentaram pouquíssima ou nenhuma relação com a variável resposta.

Foi possível notar que as variáveis explicativas então bem correlacionadas entre si. Isto pode ser um problema quando for realizado um modelo de regressão linear múltiplo, podendo gerar multicolinearidade. As variáveis expectativa de vida masculina e feminina, por exemplo, obtiveram um coeficiente de correlação de *Pearson* igual a 0,95, o que significa que elas são quase proporcionais.

Dessa forma, através da análise dos modelos de regressão linear simples, foi possível identificar que a expectativa de vida masculina junto com as despesas com educação são as variáveis que mais influenciam positivamente na renda mensal, seguida pela expectativa de vida feminina e pelo QI. Já a taxa de natalidade foi identificada como a variável que mais influencia negativamente na renda mensal.

No entanto, foi observado que algumas variáveis possuem uma relação exponencial com a renda mensal, tais como: QI, expectativa de vida masculina, expectativa de vida feminina e taxa de natalidade. Sendo necessário considerar a transformação logarítmica da renda mensal, que ajudou a reduzir a escala dos valores e tornar a relação entre as variáveis mais linear. Como o esperado, os modelos lineares simples com essas variáveis melhoraram os seus respectivos valores do coeficiente de determinação, ou seja, melhoraram a explicabilidade do modelo após a transformação logarítmica da variável dependente.

Além disso, a análise de resíduos do modelo em que a expectativa de vida masculina é utilizada para predizer a renda mensal não apresentou violação dos pressupostos. Embora

o teste utilizado para verificar a homocedasticidade dos erros tenha apontado o contrário, os métodos gráficos foram o suficiente para validar a adequabilidade desse modelo.

Portanto, a análise dos modelos apresentados pode ser utilizada como base para entender a relação entre as variáveis e a renda mensal em um nível mundial e para a construção de novos modelos que levem em consideração essas variáveis conjuntamente. Esperamos que ao utilizar o modelo de regressão linear múltiplo consigamos uma explicabilidade maior que o melhor modelo de regressão linear simples. Devido ao resultado com a transformação logarítmica, é esperado que esta transformação também seja realizada no modelo de regressão linear múltiplo, embora nem todas as variáveis precisem dessa transformação. E, será preciso lidar com a iminente multicolinearidade que aparecerá no modelo de regressão linear múltiplo.

REFERÊNCIAS

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.