



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**

Relatório apresentado 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

LISTA DE FIGURAS

Figura 3.1	<i>Boxplots</i> das variáveis socioeconômicas, demográficas e climáticas dos países em 2019	16
Figura 3.2	<i>Boxplots</i> , de acordo com o continente, das variáveis socioeconômicas, demográficas e climáticas dos países em 2019	17
Figura 3.3	Gráfico de dispersão, densidade e correlação das variáveis socioeconômicas, demográficas e climáticas dos países em 2019	18
Figura 3.4	Log-verossimilhanças do parâmetro da transformação de potência Box-Cox para as variáveis socioeconômicas, demográficas e climáticas em vários países em 2019	19
Figura 3.5	Gráfico de regressão parcial da variável natalidade em relação às variáveis socioeconômicas, demográficas e climáticas em vários países em 2019	20
Figura 3.6	Resíduos versus valores preditos e resíduos versus quantis teóricos, para o modelo sem interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019	23
Figura 3.7	Resíduos versus valores preditos e resíduos versus quantis teóricos, para o modelo com interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019	23
Figura A.1	Gráfico de dispersão, densidade e correlação das variáveis socioeconômicas, demográficas e climáticas dos países em 2019 de acordo com o continente	31

Figura B.1 Gráfico da distância de <i>cook</i> e resíduos estudentizados versus <i>leverage</i> da regressão linear em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal em vários países em 2019	32
Figura B.2 Gráfico dos DFFITS e COVARATIO da regressão linear em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal em vários países em 2019	33
Figura B.3 Gráfico dos DFBETAS da regressão linear em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal em vários países em 2019	33

LISTA DE TABELAS

Tabela 2.1	Variáveis socioeconômicas, demográficas e climáticas em diversos países em 2019.	5
Tabela 2.2	Tabela ANOVA	7
Tabela 3.1	Sumarização das variáveis socioeconômicas, demográficas e climáticas dos países em 2019	15
Tabela 3.2	Modelo de regressão, selecionado através do método <i>backward</i> baseado no teste F parcial, em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019	19
Tabela 3.3	Modelo de regressão, selecionado realizando todas as combinações possíveis baseado no AIC e coeficiente de determinação ajustado, em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019	20
Tabela 3.4	Modelo de regressão, com interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019	21
Tabela 3.5	Valores p dos testes para as suposições de normalidade, homoscedasticidade e independência dos resíduos nos modelos, sem e com interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019	22
Tabela 3.6	Coeficiente de determinação e valores VIF para as variáveis explicativas do modelo de regressão da Tabela 3.2	24
Tabela 3.7	Modelo de regressão, selecionado através do método <i>backward</i> baseado no teste F parcial, em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019	25

LISTA DE TABELAS

Tabela 3.8	Valores preditos de renda mensal para diferentes países	26
------------	---	----

SUMÁRIO

1	INTRODUÇÃO	1
2	MATERIAIS E MÉTODOS	4
2.1	Descrição dos dados	5
2.2	Modelo de regressão linear	6
2.3	Testes de hipóteses para os parâmetros	7
2.4	Coeficiente de determinação	8
2.5	Análise dos resíduos	9
2.5.1	Normalidade	9
2.5.2	Homocedasticidade	9
2.5.3	Independência	10
2.6	Transformações	10
2.7	Seleção de modelos	11
2.7.1	Teste F parcial	11
2.7.2	Critério de Informação de Akaike (AIC)	12
2.8	Seleção de variáveis	12
2.9	Observações não usuais	13
2.10	Multicolinearidade	14
3	RESULTADOS	15
3.1	Análise descritiva	15
3.2	Transformação	18
3.3	Seleção de modelos	19
3.3.1	Verificação de interação no modelo	21

SUMÁRIO

3.4	Análise dos resíduos	22
3.5	Análise de observações não usuais e multicolinearidade	23
3.6	Interpretação do Modelo	25
4	CONCLUSÃO	27
	REFERÊNCIAS	30
	Apêndice A Gráfico de dispersão das variáveis de acordo com o continente	31
	Apêndice B Gráficos para identificação de observações não usuais	32
	Apêndice C Códigos	34

1 Introdução

A desigualdade de renda é um fenômeno que ocorre quando existe uma distribuição desigual dos recursos financeiros entre os indivíduos de uma sociedade. Existem várias razões pelas quais a desigualdade de renda ocorre, como diferenças na educação, discriminação de gênero ou raça, acesso desigual a oportunidades econômicas, concentração de poder econômico e político nas mãos de poucos, políticas governamentais inadequadas e falta de mobilidade social.

A desigualdade de renda é um problema persistente em muitas partes do mundo, e sua redução é um dos principais desafios enfrentados pelos governos e organizações internacionais. De acordo com Chancel *et al.* (2022), os 10% mais ricos do mundo ganham 52% da renda mundial, enquanto os 50% mais pobres recebem apenas 8,5% do total. Dessa forma, fica evidente que a concentração de renda é um problema global que exige atenção.

Os efeitos da desigualdade de renda são vastos e afetam diferentes aspectos da sociedade. Ela pode levar a uma diminuição da coesão social, disparidades no acesso a serviços básicos, aumentar a criminalidade e a instabilidade política. Ela também pode limitar as oportunidades de desenvolvimento econômico e reduzir o crescimento sustentável a longo prazo.

Dessa forma, a redução desse problema é fundamental para combater os efeitos adversos causados por disparidades socioeconômicas. De acordo com Barros (2006), o aumento acelerado na renda média dos mais pobres reduz a pobreza. A queda na pobreza, por sua vez, resulta tanto do crescimento econômico balanceado, que eleva igualmente a renda de todos os grupos, quanto das reduções no grau de desigualdade, as quais elevam a fatia dos pobres na renda nacional e reduzem a dos ricos. Assim, através de estratégias combinadas, é possível construir uma sociedade mais justa e inclusiva, onde as oportunidades sejam equitativamente distribuídas e os efeitos negativos da desigualdade sejam mitigados.

Por isso, compreender os fatores que influenciam a renda mensal é fundamental para desenvolver estratégias eficazes de redução da desigualdade e promoção de um desenvolvimento econômico mais equitativo. Este projeto de pesquisa tem como objetivo investigar a relação entre a renda mensal e um conjunto de variáveis socioeconômicas,

demográficas e climáticas em vários países. Para tanto, será realizada uma análise de regressão linear múltipla, que permite examinar como essas variáveis explicativas estão relacionadas à renda mensal. Os objetivos específicos dessa análise incluem:

- Realizar uma análise descritiva dos dados para compreender as características gerais da renda e das variáveis socioeconômicas, demográficas e climáticas.
- Quantificar a relação entre a renda mensal e as variáveis explicativas, por meio do modelo de regressão linear, identificando os fatores que apresentam maior influência.
- Discutir os padrões e tendências encontrados.

As variáveis explicativas consideradas incluem a expectativa de vida masculina e feminina, taxa de mortalidade, taxa de natalidade, índice de inteligência, gastos com educação por habitante, temperatura máxima diária e continente.

A expectativa de vida masculina e feminina é uma medida importante que reflete a qualidade de vida. De acordo com Biciunaite (2018), pessoas nascidas em países ricos vivem mais do que em nações pobres. Países com renda mais alta geralmente podem ter uma expectativa de vida maior devido ao acesso a melhores cuidados de saúde, alimentação adequada, educação e condições de vida melhores. A inclusão dessas variáveis permite investigar como a longevidade está relacionada à renda mensal.

Em relação a taxa de mortalidade, a renda mais baixa está relacionada a uma maior taxa de mortalidade, devido à falta de acesso a serviços de saúde, moradia precária, alimentação inadequada e exposição a ambientes perigosos. Já taxa de natalidade, segundo Nargund (2009) em países com renda mais alta, as taxas tendem a ser mais baixas devido aos níveis mais altos de educação e métodos contraceptivos. As taxas de mortalidade e natalidade influenciam a composição demográfica e podem impactar diretamente a renda da população.

Ademais, considerar o índice de inteligência e os gastos com educação por habitante permite entender como o capital intelectual e os investimentos educacionais podem influenciar a renda mensal. Países que valorizam a educação e promovem o desenvolvimento de habilidades e conhecimentos têm maior probabilidade de obter uma força de trabalho qualificada e competitiva, impulsionando a renda.

As variáveis climáticas, como a temperatura máxima diária, também podem desempenhar um papel importante na determinação da renda mensal. Climas extremos, como altas temperaturas, podem afetar a produtividade agrícola e industrial, levando a variações na renda em diferentes regiões. Incluir essa variável na análise permite avaliar como o clima influencia a renda mensal e quais regiões podem ser mais vulneráveis a esses efeitos.

A variável continente é adicionada para capturar as diferenças que podem existir entre os países de diferentes regiões geográficas. Cada continente pode ter características

socioeconômicas, políticas e culturais distintas que afetam a distribuição de renda e as oportunidades econômicas disponíveis. Ao considerar essa variável, é possível examinar como a localização geográfica influencia a renda mensal.

Assim sendo, este estudo fornecerá uma compreensão mais abrangente das relações entre a renda mensal e as variáveis socioeconômicas, demográficas e climáticas consideradas.

2 Materiais e métodos

Os métodos descritos nesta seção serão aplicados aos dados de diversos países. Todas as análises serão conduzidas utilizando o software R (R Core Team, 2022). Inicialmente, será realizada uma análise da relação entre a renda mensal e as variáveis explicativas. Caso essa relação não seja linear, será proposta uma transformação da renda mensal com base na transformação Box-Cox.

Em seguida, serão utilizados modelos de regressão linear múltipla para modelar a relação da renda mensal com as variáveis explicativas selecionadas. Será feito uso do método de seleção de variáveis *backward*, além de considerar todas as regressões possíveis com base nos critérios AIC e teste F parcial. A significância global e individual dos parâmetros será observada, bem como a proporção da variabilidade original dos dados explicada pelos modelos.

Serão utilizados gráficos da densidade empírica dos resíduos, qq-plot e o teste de normalidade de Shapiro-Wilk para avaliar a normalidade dos resíduos. Serão examinados gráficos de dispersão dos resíduos em relação aos valores ajustados e será aplicado o teste de Goldfeld-Quandt para verificar a homocedasticidade. Também será utilizado o teste de Durbin-Watson para verificar a independência dos erros.

Além disso, a identificação e avaliação das observações não usuais por meio do leverage, DFBetas, DFFits, distância de Cook e COVRATIO que são de extrema importância para a análise dos dados. Essas medidas permitem identificar observações que possuem um impacto significativo nos resultados do modelo de regressão, seja por sua influência nos coeficientes estimados, nos valores ajustados ou na variabilidade dos resíduos. Por fim, a presença de multicolinearidade será avaliada utilizando o fator de inflação da variância (VIF) e a tolerância.

2.1 Descrição dos dados

A base de dados contém informações socioeconômicas, demográficas e climáticas de 85 países, disponíveis no site <https://www.worlddata.info/>. Os dados foram obtidos através da combinação de bases de dados contendo informações sobre a expectativa de vida, quociente de inteligência e custo de vida de cada país em 2019.

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 socioeconômicas, demográficas e climáticas em diversos países em 2019.

Essas variáveis foram selecionadas por sua relevância na análise socioeconômica, demográfica e climática dos países.

2.2 Modelo de regressão linear

O modelo de regressão linear é uma técnica utilizada para analisar e modelar a relação entre uma variável resposta contínua e uma ou mais variáveis explicativas. De acordo com Demétrio e Zocchi (2006), o modelo de uma regressão linear com p variáveis explicativas é:

$$\mathbf{Y}_{(nx1)} = \mathbf{X}_{(nxk)}\boldsymbol{\beta}_{(kx1)} + \boldsymbol{\epsilon}_{(nx1)} \quad (2.1)$$

em que $k = p + 1$; \mathbf{Y} é o vetor da variável resposta; \mathbf{X} é chamada de matriz de delineamento, ela contém as observações das variáveis explicativas, bem como uma coluna adicional de 1's, para representar o coeficiente β_0 ; $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos; $\boldsymbol{\epsilon}$ é o vetor de variáveis aleatórias não observáveis (erros). Dessa forma, têm-se as suposições:

- i. as variáveis explicativas são fixas;
- ii. a variável resposta é função linear das variáveis explicativas, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, sendo $\mathbf{0}$ um vetor de zeros de dimensões $nx1$;
- iii. os erros são homocedásticos, $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$, sendo \mathbf{I} a matriz identidade;
- iv. os erros são independentes, $\text{Cov}(\epsilon_i, \epsilon_j) = E(\epsilon_i, \epsilon_j) = 0$, $i \neq j$;
- v. os erros têm distribuição normal, $\epsilon_i \sim N(0, \sigma^2)$.

O método de mínimos quadrados é utilizado para estimar os parâmetros do modelo, ele é utilizado pois possui ótimas propriedades: ele é o estimador não viciado de variância uniformemente mínima de $\boldsymbol{\beta}$.

O método consiste em estimar os parâmetros do modelo de tal forma que os desvios dos valores observados em relação aos estimados sejam mínimos, isso equivale a minimizar o comprimento do vetor $\boldsymbol{\epsilon} = \mathbf{Y} - E(\mathbf{Y})$. Usando a norma euclidiana para avaliar com comprimento desse vetor, tem-se:

$$z = \|\boldsymbol{\epsilon}\|^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = \sum_i^n \epsilon_i^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^T \mathbf{Y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \boldsymbol{\beta} \mathbf{X}.$$

Como matriz de delineamento tem posto coluna completo, então o sistema de equações normais $\partial z / \partial \boldsymbol{\beta} = 0$ é consistente e tem solução única dada por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

Assim, para compreender o impacto das variáveis explicativas na variável resposta é possível interpretar dos parâmetros estimados. O intercepto do modelo, representa o valor

esperado da variável resposta quando todas as variáveis explicativas são zero. Em termos práticos, indica o valor inicial da variável resposta quando as outras variáveis explicativas não têm efeito. Os coeficiente angular representa a mudança esperada na variável resposta para uma unidade de aumento em X_j , mantendo todas as outras variáveis explicativas constantes. Se $\hat{\beta}_j$ for positivo, isso indica uma relação positiva entre X_j e a variável resposta. Se for negativo, indica uma relação negativa.

2.3 Testes de hipóteses para os parâmetros

Para testar a significância conjunta dos parâmetros do modelo, isto é, testar $H_0 : \beta_1 = \dots = \beta_p = 0$ versus H_1 : pelo menos um β_j difere de zero, é possível utilizar a análise de variância (ANOVA). Com base em Neter *et al.* (1996), as somas de quadrados são dadas por

$$\begin{aligned} SQ_{Total} &= \mathbf{Y}^T \mathbf{Y} - \frac{1}{n} \mathbf{Y}^T \mathbf{J} \mathbf{Y}, \\ SQ_{Reg} &= \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{Y} - \frac{1}{n} \mathbf{Y}^T \mathbf{J} \mathbf{Y} \quad \text{e} \\ SQ_{Res} &= \mathbf{Y}^T \mathbf{Y} - \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{Y}, \end{aligned}$$

em que \mathbf{J} é uma matriz $n \times n$ de 1's. A soma de quadrados totais tem $n - 1$, a soma de quadrados da regressão tem $k - 1$ e a soma de quadrados dos resíduos tem $n - k$ graus de liberdade associados. A Tabela 2.2 mostra esses resultados de análise de variância, bem como os quadrados médios.

Tabela 2.2: Tabela ANOVA

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	Estatística F
Regressão	$k - 1$	SQReg	$QMReg = \frac{SQReg}{k-1}$	$\frac{QMReg}{QMRes}$
Resíduo	$n - k$	SQRes	$QMRes = \frac{SQRes}{n-k}$	
Total	$n - 1$	SQTotal		

Sob a hipótese nula, a estatística F segue uma distribuição F de Snedecor com $k - 1$ e $n - k$ graus de liberdade. Se o valor-p associado com a estatística F for menor que o nível de significância escolhido, rejeitamos a hipótese nula e concluímos que pelo menos um dos coeficientes é significativo. Caso contrário, não rejeitamos a hipótese nula e concluímos que não há evidências suficientes para concluir que os coeficientes são significativos.

Para testar a significância individual dos parâmetros do modelo, o teste t-Student pode ser utilizado. Assim, com base em Hoffmann (2016), para testar $H_0 : \beta_j = 0$ versus $\beta_j \neq 0$, isto é, testar a significância de β_j no modelo, utiliza-se a estatística de teste dada

por

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{var}(\beta_j)}},$$

em que $\text{var}(\hat{\beta}_j) = \hat{\sigma}^2(\mathbf{X}^T \mathbf{X})_{jj}^{-1}$, sendo $\hat{\sigma}^2 = QMRes$ e $(\mathbf{X}^T \mathbf{X})_{jj}^{-1}$ o j -ésimo termo da diagonal principal de $(\mathbf{X}^T \mathbf{X})^{-1}$.

Sob a hipótese nula, essa estatística de teste segue uma distribuição t-Student com $n - k$ graus de liberdade. Se o valor-p associado com a estatística t for menor que o nível de significância escolhido, rejeitamos a hipótese nula e concluimos que o coeficiente β_j é significativo. Caso contrário, não rejeitamos a hipótese nula e concluimos que não há evidências suficientes para concluir que o coeficiente é significativo.

2.4 Coeficiente de determinação

O coeficiente de determinação expressa a proporção da variabilidade original dos dados explicada pelo modelo de regressão. De acordo com Neter *et al.* (1996), ele é denotado por R^2 e definido como:

$$R^2 = \frac{SQReg}{SQTtotal} = 1 - \frac{SQRes}{SQTtotal},$$

em que $0 \leq R^2 \leq 1$. Onde R^2 assume o valor 0 quando todos os $\beta_j = 0$, e o valor 1 quando todas as observações da variável resposta caem diretamente na superfície de regressão ajustada, i.e., quando $Y_j = \hat{Y}_j \forall j$.

Como R^2 geralmente pode ser aumentado incluindo um maior número de variáveis explicativas, às vezes é sugerido que uma medida modificada seja utilizada. O coeficiente ajustado, denotado por R_a^2 , ajusta R^2 dividindo cada soma de quadrados por seus graus de liberdade associado:

$$R_a^2 = 1 - \frac{SQRes/(n - k)}{SQTtotal/(n - 1)}.$$

Esse coeficiente ajustado pode se tornar menor quando outra variável explicativa é introduzida no modelo, porque qualquer diminuição na $SQRes$ pode ser mais do que compensada pela perda de um grau de liberdade no denominador $n - k$.

2.5 Análise dos resíduos

Para que os resultados do modelo sejam válidos, é necessário que sejam atendidas algumas suposições importantes: normalidade, homogeneidade de variâncias (homocedasticidade) e independência.

2.5.1 Normalidade

É importante que as distribuições de cada grupo sejam aproximadamente normais. Existem diferentes métodos para avaliar a normalidade dos dados, como o gráfico de probabilidade normal (*QQ plot*) e o teste de *Shapiro-Wilk*.

O teste proposto por Shapiro e Wilk (1965) é um procedimento alternativo para avaliar normalidade. A estatística de teste é dada por

$$W = \frac{b^2}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

em que $b_2 = \frac{n}{2} \sum_{i=1}^n a_{n-i+1} (X_{n-i+1} - X_i)$ e a_j são valores obtidos nas tabelas de *Shapiro-Wilks*. Para este procedimento, ordenam-se as n observações da amostra, tal que: $X_1 \leq X_2 \leq \dots \leq X_n$.

A estatística de teste W pode representar a correlação entre os dados e seus correspondentes escores normais. Assim, quando $W = 1$ os dados ajustam-se perfeitamente à distribuição normal. Valores críticos do teste foram obtidos por simulação e também encontram-se tabelados. Assim, rejeita-se a hipótese de normalidade quando $W_{\text{calc}} > W_{\text{crit}}$.

2.5.2 Homocedasticidade

A homocedasticidade refere-se à variabilidade dos dados dentro de cada grupo. A homogeneidade de variâncias pode ser verificada visualmente, por meio do gráfico de dispersão dos dados, ou por meio de testes, como o teste de *Breush-Pagan*.

O teste de *Breush-Pagan* (Breusch e Pagan, 1979), é baseado no ajuste de um modelo de regressão em que a variável dependente é definida pelos resíduos do modelo de interesse.

Se grande parte da variabilidade dos resíduos é explicada pelo modelo, então rejeita-se a hipótese de homocedasticidade. A estatística de teste é dada por

$$BP_{\text{calc}} = nR^2,$$

que segue uma distribuição qui-quadrado com um grau de liberdade. Assim, rejeita-se a

hipótese de homocedasticidade quando $BP_{calc} > \chi^2_{1crit}$.

2.5.3 Independência

A independência refere-se ao fato de que as observações em cada grupo devem ser independentes entre si. Testes estatísticos podem ser utilizados para verificar a independência, como por exemplo, o teste de *Durbin-Watson*, proposto por Durbin e Watson (1992). O teste tem estatística de teste definida em termos dos resíduos, tal que

$$d = \frac{\sum_{i=2}^n (e_t - e_{n-1})^2}{\sum_{i=1}^n e_n^2}.$$

O valor da estatística d encontra-se entre 0 e 4. A avaliação descritiva dos valores de d pode ser feita da seguinte forma:

- $d = 2$ não há autocorrelação;
- $d < 1$ há evidência de correlação serial positiva;
- $d > 3$ há evidência de correlação serial negativa.

Como regra geral se $1,5 < d < 2,5$ indica independência das observações e a má especificação do modelo de regressão ou a ausência de variáveis importantes no modelo podem resultar em autocorrelação dos erros.

2.6 Transformações

Se o modelo de regressão não for apropriado para um conjunto de dados, ou seja, violar algumas das suposições, existem duas escolhas básicas:

1. Abandonar o modelo de regressão (2.1) e desenvolver e usar um modelo mais apropriado.
2. Empregar alguma transformação nos dados para que o modelo de regressão (2.1) seja apropriado para os dados transformados.

Cada abordagem tem vantagens e desvantagens. A primeira abordagem pode implicar um modelo mais complexo que pode render melhores percepções, mas também pode levar a procedimentos mais complexos para estimar os parâmetros. O uso bem-sucedido de transformações, por outro lado, leva a métodos de estimativa relativamente simples e

pode envolver menos parâmetros do que um modelo complexo, uma vantagem quando o tamanho da amostra é pequeno.

Existem varias transformações que podem ser aplicadas para aproximar a distribuição dos dados da normalidade e homocedasticidade. De acordo com Demétrio e Zocchi (2006), as transformações mais utilizadas são: logarítmica ($\log(Y)$), raiz quadrada (\sqrt{Y}), reciprocidade ($1/Y$), quadrática (Y^2) e Box-Cox.

A transformação de Box-Cox (Box e Cox, 1964) é uma transformação mais geral que pode ser aplicada para corrigir diferentes violações dos pressupostos. A família de transformações é dada por

$$Y(\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log(Y) & \text{se } \lambda = 0. \end{cases}$$

sendo λ o parâmetro da transformação e Y a variável resposta, na ausência de uma transformação, $\lambda = 1$. O λ varia continuamente, permitindo diferentes transformações, incluindo as transformações logarítmica, de raiz quadrada, de reciprocidade e quadrática.

2.7 Seleção de modelos

Usar critérios que auxiliem na comparação de modelos paramétricos é de fundamental importância. Para fazer isso, é possível utilizar o teste F parcial e o Critério de Informação de Akaike (AIC).

2.7.1 Teste F parcial

A comparação entre o modelo completo e o modelo reduzido pode ser feita utilizando o teste F parcial. Considerando que a hipótese de interesse é $H_0 : \beta_1 = \beta_2 = \dots = \beta_{q*} = 0$, $q* < p$, a estatística de teste é dada por:

$$F_o = \frac{(SQReg_c - SQReg_r)/(k - q - 1)}{(SQRes_c)/(n - k)}$$

em que q é o número de parâmetros do modelo restrito -1 e $k = p + 1$ é o número de parâmetros estimados no modelo. Sob a hipótese nula, essa estatística segue uma distribuição F com $k - q - 1$ e $n - k$ graus de liberdade. Se o valor de F_o for maior que o valor crítico da distribuição F para um determinado nível de significância, a hipótese nula é rejeitada e concluímos a favor do modelo completo.

2.7.2 Critério de Informação de Akaike (AIC)

O Critério de Informação de Akaike (Akaike, 1974), é baseado na medida de de Informação de Kullback-Leibler. Ele é obtido por

$$AIC = -2 \log L + 2k,$$

em que k é o número de parâmetros do modelo, n é o número de observações e L é a função de verossimilhança. No caso do modelo de regressão linear, temos:

$$AIC = -n \log(SQRes/n) + 2k.$$

Dado um conjunto de modelos candidatos, o preferido será o que fornecer o menor AIC. Além de selecionar um ótimo ajuste, o critério penaliza a adição de parâmetros, desencorajando a seleção de um modelo extremamente complexo e com muitos parâmetros que tenham um pobre desempenho.

2.8 Seleção de variáveis

A seleção de covariáveis é essencial para obter um modelo simples e explicativo. Ela exclui variáveis desnecessárias que podem influenciar a estimação dos parâmetros e desperdiçar graus de liberdade. Além disso, reduz a colinearidade entre as covariáveis e pode resultar em economia de custos. A seleção busca encontrar um equilíbrio entre relevância e parcimônia, mantendo as variáveis mais importantes e descartando as redundantes. Isso melhora a interpretabilidade e a eficiência do modelo, proporcionando um ajuste preciso ao fenômeno de interesse.

Existem diferentes métodos para selecionar variáveis, mas os métodos utilizados neste relatório são o *backward* e o método de todas as regressões possíveis.

O método de todas as regressões possíveis envolve o ajuste de 2^k modelos. Os modelos, são separados em grupos de modelos com b variáveis, $b = 1, 2, \dots, k$ sendo cada grupo ordenado de acordo com algum critério, por exemplo, AIC, escolhendo-se os modelos com menor AIC de cada grupo ou, então, menor QMRes. Se os dois modelos apresentam valores próximos de AIC e de QMRes, escolhe-se aquele com menor número de parâmetros.

O método *backward*, baseado no AIC por exemplo, começa com um modelo que inclui todas as covariáveis consideradas inicialmente relevantes. Em seguida, itera-se removendo uma covariável por vez, calculando o valor do AIC para cada novo modelo resultante. A covariável menos relevante, aquela cuja exclusão leva ao menor aumento no valor do AIC,

é então removida do modelo. Esse processo é repetido até que não seja possível remover mais covariáveis sem que o valor do AIC aumente significativamente.

2.9 Observações não usuais

A identificação de observações não usuais é uma parte importante da modelagem de dados. Essas observações são pontos que diferem significativamente dos demais e podem exercer uma influência desproporcional nos resultados da análise.

As observações não usuais podem ser classificadas em três tipos: *outliers* que são observações que não são bem ajustadas pelo modelo; observações influentes que são observações que afetam alguma propriedade do modelo ajustado de maneira substancial; e ponto de alavanca que é um ponto extremo no espaço das variáveis explicativas.

A forma mais eficiente de detectar observações não usuais é através da matriz chapéu:

$$H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

e h_{ii} (leverage) os valores da diagonal principal de H . Pontos de alavanca apresentam h_{ii} grande e *outliers* apresentam h_{ii} pequeno.

Além disso, de acordo com Fox (1991) vários métodos têm sido propostos para medir a influência:

- DFBetas: medem a influência de cada observação nos coeficientes de regressão do modelo. São calculados como a diferença entre o valor estimado do coeficiente com e sem a observação em questão, dividida pelo erro padrão estimado do coeficiente. DFBetas maiores do que $2/\sqrt{n}$ (ou menores do que $-2/\sqrt{n}$) indicam uma influência considerável da observação correspondente no coeficiente de regressão.
- DFFITS: medem a influência global de cada observação nos valores ajustados (ou preditos) do modelo. Essa medida quantifica a diferença entre os valores ajustados com e sem a observação em questão, normalizada pelo erro padrão dos resíduos. DFFITS maiores do que $2/\sqrt{k/n}$ (ou menores do que $-2/\sqrt{k/n}$) indicam uma influência substancial da observação nos valores ajustados.
- Distância de Cook: avalia a influência global de uma observação, considerando tanto os coeficientes de regressão quanto os valores ajustados. Ela é calculada como uma combinação ponderada dos DFBetas e dos DFFITS para cada observação. Uma distância de Cook maior do que $4/n$ indica uma influência significativa da observação no modelo.

- COVRATIO: é uma medida usada para detectar observações influentes que afetam a variância dos resíduos do modelo. Ele compara a variância dos resíduos estimada com e sem a observação em questão. Um COVRATIO maior do que $3k/(n-k)$ indica que a observação tem uma influência desproporcional na variância dos resíduos.

2.10 Multicolinearidade

A multicolinearidade se caracteriza por uma quase dependência linear entre as variáveis explicativas. Essa correlação levar a imprecisão das estimativas dos coeficientes da regressão, impacto nos testes de hipóteses e interpretação das estimativas e instabilidade computacional.

A multicolinearidade pode ser detectada através de várias métricas, como o coeficiente de correlação, a matriz de correlação ou o fator de inflação da variância (VIF). O VIF indica a velocidade com a qual as variâncias e covariâncias aumentam, um VIF elevado indica alta multicolinearidade entre uma variável independente e as demais. De acordo Gujarati e Porter (2011), o VIF é definido por:

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, \dots, p$$

em que R_j^2 é o coeficiente de determinação da regressão de X_j nas demais variáveis. O VIF mostra como a variância de um estimador é inflada pela presença da multicolinearidade. Quando R^2 aproxima-se de 1, o VIF aproxima-se do infinito. Ou seja, quando a colinearidade aumenta, a variância de um estimador aumenta e, no limite, pode tornar-se infinita. Se não houver colinearidade entre X_i e X_j , o VIF será 1. Uma regra prática para indicação de multicolinearidade é a identificação de qualquer $VIF_j > 10$

O inverso de VIF é chamado de tolerância (TOL),

$$TOL_j = \frac{1}{VIF_j} = (1 - R_j^2)$$

Quando $R_j^2 = 1$ (colinearidade perfeita), $TOL_j = 0$ e $R_j^2 = 0$ (não há colinearidade nenhuma), $TOL_j = 1$. Devido à ligação estreita entre FIV e TOL, eles podem ser usados indistintamente.

3 Resultados

Nesta seção serão apresentados os resultados da análise descritiva dos dados, das transformações realizadas e da modelagem dos dados.

Um ponto a se ressaltar é que o continente Oceania foi utilizado nas análises descritivas como um continente à parte. Todavia, devido a ter apenas dois países no continente, optou-se por incluir a Oceania no continente da Ásia, já que há proximidade geográfica entre os dois continentes para os modelos de regressão linear múltiplo.

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, já que apresenta um coeficiente de variação acima de 1, o que é resultado da grande diferença de renda e de gastos com a educação entre os países.

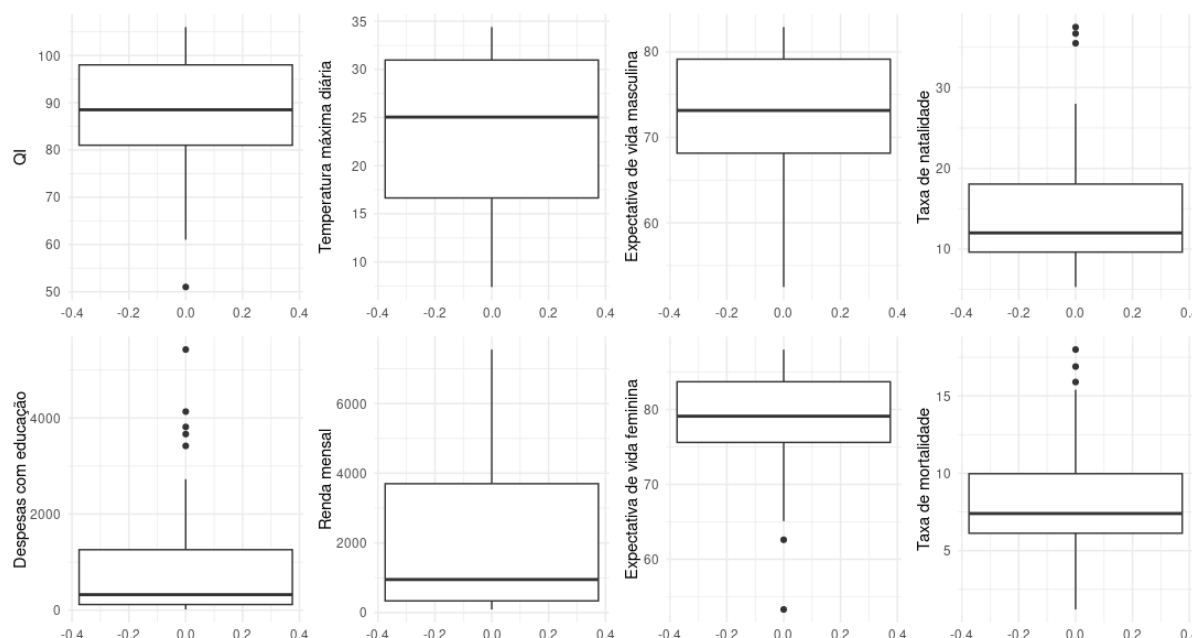
Tabela 3.1: Sumarização das variáveis socioeconômicas, demográficas e climáticas dos países em 2019

	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
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
Máximo	106,0	5425,0	34,4	7550,0	82,9	88,0	37,5	18,0
CV	0,12	1,27	0,35	1,03	0,08	0,07	0,47	0,41

Observando os *boxplots* da Figura 3.1, é possível notar que essa grande variabilidade da renda mensal e dos gastos com educação 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 3.1: *Boxplots* das variáveis socioeconômicas, demográficas e climáticas dos países em 2019



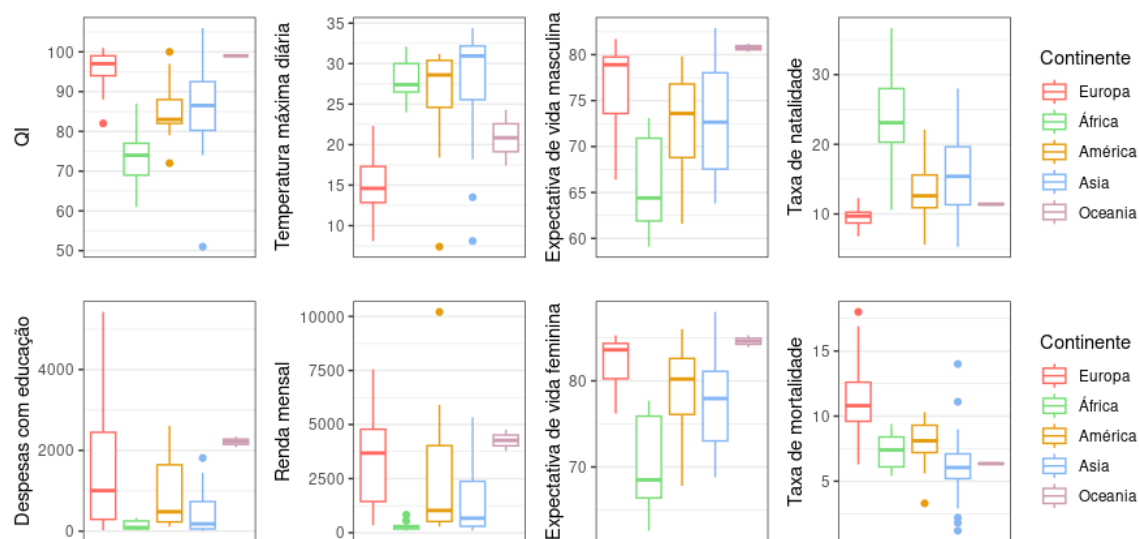
A Oceania tem apenas dois países: Austrália e Nova Zelândia, portanto é mais complicado analisar as características de dispersão das variáveis em relação a esse continente. Mas, é possível observar na Figura 3.2, que em relação ao QI, a Oceania apresenta a maior mediana, seguida por Europa, enquanto 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 à 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 e maiores valores na Europa e com uma baixa variabilidade e menores valores na África e na América, sendo que a Oceania

Figura 3.2: *Boxplots*, de acordo com o continente, das variáveis socioeconômicas, demográficas e climáticas dos países em 2019



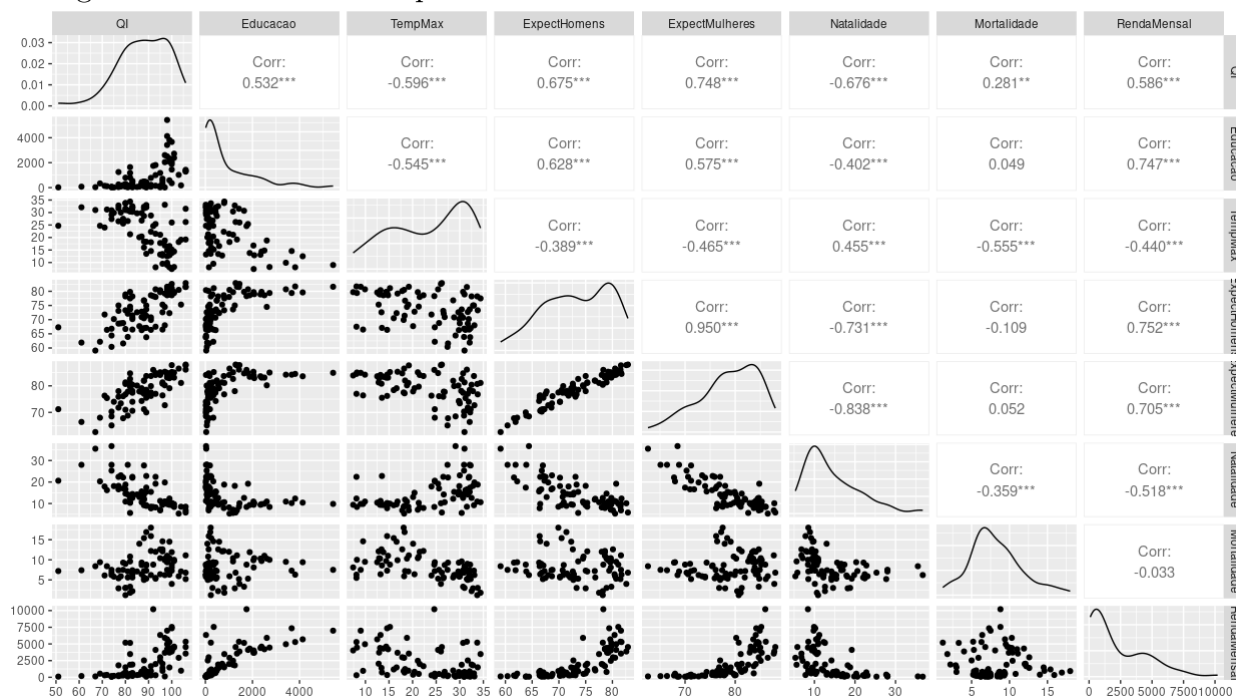
apresenta a maior mediana para essas variáveis. A América apresenta valores de mediana das despesas com educação e da renda mensal maiores do que a mediana das da Ásia, além também de apresentar uma maior dispersão.

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.

Observando a Figura 3.3, as variáveis no geral apresentam correlações significativas entre si. Entretanto, parece haver diferenças nos coeficientes de correlações lineares condicionados pelo continente. Por exemplo, na Europa, a correlação entre mortalidade e despesas com educação é igual a $-0,638$, ou seja, a uma correlação moderada negativa entre as duas variáveis. Por outro lado, na África, a correlação linear entre essas duas variáveis é igual a $0,364$, indicando uma correlação moderada positiva, sendo a correlação geral dessa variável $0,049$ (correlação fraca). Portanto, o continente parece afetar o modo como as variáveis se relacionam.

Quando as variáveis explicativas tem um coeficiente de correlação de *Pearson* com valor absoluto alto com a variável resposta, isto é algo positivo para o modelo de regressão linear. Mas, quando as variáveis independentes apresentam uma correlação absoluta alta entre si, pode haver problema de multicolinearidade no modelo de regressão.

Figura 3.3: Gráfico de dispersão, densidade e correlação das variáveis socioeconômicas, demográficas e climáticas dos países em 2019



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 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.

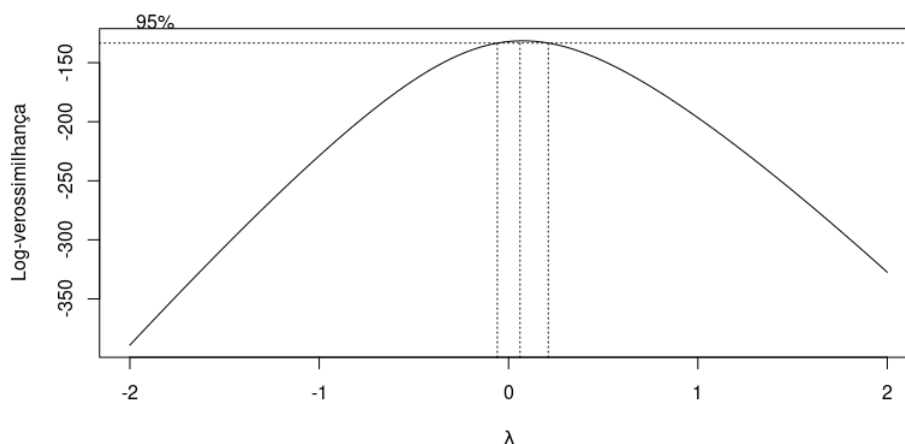
3.2 Transformação

Devido ao que foi observado na análise descritiva sobre a relação exponencial que certas variáveis têm com a renda mensal, foi proposto realizar a transformação Box-Cox na variável resposta.

Na família proposta por Box-Cox, a transformação mais adequada é aquela que utiliza um parâmetro de transformação (lambda) próximo a 0,061, conforme mostrado na Figura

8. No entanto, como é possível observar na Figura 3.4, o intervalo de confiança indica que a transformação logarítmica é uma opção viável, uma vez que o intervalo de confiança inclui o valor zero. Desse modo, esta transformação será aplicada na variável resposta.

Figura 3.4: Log-verossimilhanças do parâmetro da transformação de potência Box-Cox para as variáveis socioeconômicas, demográficas e climáticas em vários países em 2019



3.3 Seleção de modelos

Será utilizado o teste F parcial, o AIC e o coeficiente de determinação ajustado para selecionar as variáveis mais relevantes para prever o logaritmo da renda mensal dentro do modelo de regressão linear múltiplo. A seleção de variáveis *backward* baseada no teste F parcial forneceu os resultados presentes na Tabela 3.2. Por outro lado, realizando todas as combinações possíveis de modelos com essas variáveis, o que obteve o menor AIC e menor coeficiente de determinação foi o modelo da Tabela 3.3.

Tabela 3.2: Modelo de regressão, selecionado através do método *backward* baseado no teste F parcial, em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019

	Coeficiente	EP	Est. T	Valor-p
Intercepto	-4,768	1,049	-4,544	<0,001
QI	0,020	0,009	2,394	0,019
Educacao	0,001	0,0001	2,914	0,005
ExpectHomens	0,135	0,015	8,927	<0,001
Continente				
África	-0,057	0,267	-0,215	0,831
América	0,252	0,198	1,269	0,208
Asia	-0,337	0,157	-2,150	0,035

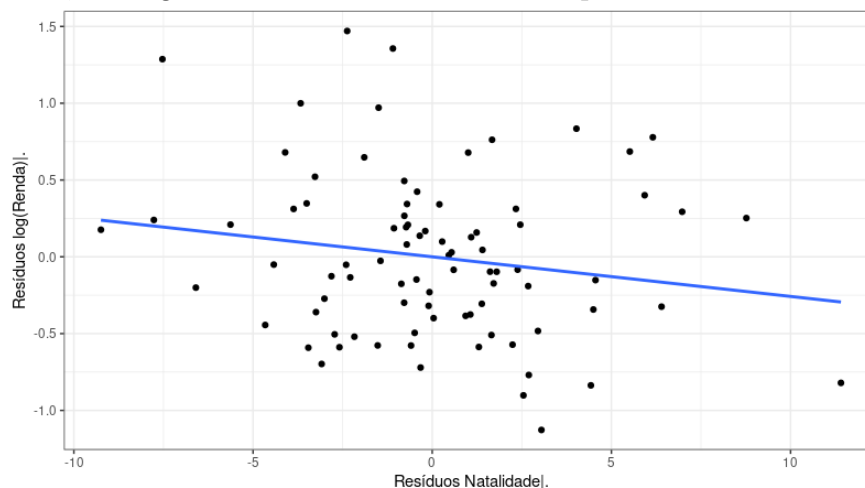
Tabela 3.3: Modelo de regressão, selecionado realizando todas as combinações possíveis baseado no AIC e coeficiente de determinação ajustado, em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal em vários países em 2019

	Coeficiente	EP	Est. T	Valor-p
Intercepto	-3,011	1,541	-1,954	0,0544
QI	0,017	0,009	1,893	0,0621
Educacao	0,0003	0,0001	3,280	0,0016
ExpectHomens	0,119	0,018	6,551	<0,001
Natalidade	-0,026	0,017	-1,545	0,1265
Continente				
África	0,133	0,292	0,456	0,6498
América	0,266	0,198	1,351	0,1808
Asia	-0,224	0,172	-1,302	0,1968

A diferença entre os dois modelos selecionados é apenas a variável natalidade. Logo, podemos através do gráfico de regressão parcial avaliar se há necessidade de incluir essa variável no modelo. Observando o gráfico dos resíduos da regressão sem a variável natalidade contra os resíduos da regressão com a variável natalidade como resposta na Figura 3.5, não é possível perceber com clareza a linearidade dessa relação. Realizando o teste t para a regressão linear simples desses dois resíduos, sendo que os resíduos da regressão sem a variável natalidade é a variável resposta, obteve-se um valor-p igual a 0,113, logo, a hipótese nula de não linearidade entre as duas variáveis não é rejeitada ao nível de 5% de confiança.

Desse modo, o modelo sumarizado na Tabela 3.2 será considerado para as análises seguintes.

Figura 3.5: Gráfico de regressão parcial da variável natalidade em relação às variáveis socioeconômicas, demográficas e climáticas em vários países em 2019



3.3.1 Verificação de interação no modelo

A inclusão da variável categórica continente se mostrou relevante, já que foi selecionada pelas três métricas utilizadas para a seleção de modelos. Mas, a variável categórica entrou no modelo apenas como efeito aditivo, ou seja, não foi considerada a interação da variável continente com as outras variáveis explicativas quantitativas. Desse modo, será realizado, novamente um teste F parcial, para comparar o modelo com efeitos aditivos com o modelo com efeitos multiplicativos. O valor-p do teste F parcial foi de aproximadamente zero, portanto, rejeita-se a hipótese nula de que o modelo de efeitos multiplicativos é equivalente ao modelo de efeitos aditivos ao nível de 5%.

Portanto, podemos considerar o modelo de efeitos multiplicativos sumarizado na Tabela 3.4 como um modelo de regressão linear possivelmente adequado para prever o logaritmo da renda mensal. Entretanto, ainda é necessário analisar os resíduos desse modelo para verificar se ele não viola as suposições de normalidade, homoscedasticidade e independência.

Tabela 3.4: Modelo de regressão, com interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019

	Coeficiente	EP	Est. T	Valor-p
Intercepto	-8,673	2,548	-3,405	0,001
QI	0,080	0,025	3,208	0,002
Educacao	0,0001	0,0001	0,469	0,640
ExpectHomens	0,115	0,025	4,519	<0,001
Continente:				
África	12,310	3,390	3,631	<0,001
América	10,050	3,681	2,731	0,008
Asia	2,457	2,970	0,827	0,411
QI x ContinenteÁfrica	-0,069	0,038	-1,821	0,073
QI x ContinenteAmérica	-0,089	0,037	-2,405	0,019
QI x ContinenteAsia	-0,072	0,026	-2,749	0,008
Educacao x ContinenteÁfrica	0,005	0,006	3,054	0,003
Educacao x ContinenteAmérica	0,001	0,0003	3,526	0,001
Educacao x ContinenteAsia	0,0003	0,0002	1,559	0,124
ExpectHomens x ContinenteÁfrica	-0,109	0,048	-2,287	0,025
ExpectHomens x ContinenteAmérica	-0,037	0,041	-0,903	0,370
ExpectHomens x ContinenteAsia	0,049	0,034	1,455	0,150

3.4 Análise dos resíduos

Realizando a análise de resíduos para o modelo sem interação e para o modelo com interação, é possível observar algumas diferenças entre os dois modelos. Na Tabela 3.5, os testes de *Shapiro-Wilk*, *Breusch-Pagan* e *Durbin-Watson* não rejeitam as suas respectivas hipóteses nulas ao nível de significância de 5% para o modelo sem interação, portanto, as suposições de normalidade, homoscedasticidade e independência, respectivamente, não foram violadas. Ao observar os gráficos dos resíduos do modelo sem interação na Figura 3.6, é notável que a variância dos resíduos aparenta ser constante ao longo dos valores preditos, apesar de três observações acima do valor 2, e que os quantis empíricos dos resíduos seguem os quantis teóricos da normal, apesar do fugir nas caudas, principalmente na cauda à direita.

Tabela 3.5: Valores p dos testes para as suposições de normalidade, homoscedasticidade e independência dos resíduos nos modelos, sem e com interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019

Modelo	Shapiro-Wilk	Breusch-Pagan	Durbin-Watson
Sem interação	0,165	0,157	0,238
Com interação	<0,001	0,144	0,405

Por outro lado, o teste de *Shapiro-Wilk* foi rejeitado ao nível de 5% de significância para o modelo com interação, logo, os resíduos estão violando a suposição de normalidade. E, ao analisar o gráfico dos resíduos versus os quantis teóricos da normal na Figura 3.7, é possível observar que os quantis empíricos dos resíduos foram bem superestimados na cauda à direita. Por outro lado, o gráfico dos resíduos versus valores preditos também na Figura 3.7 é bem similar com o mesmo gráfico para o modelo sem interação, portanto, com variância constante, o que também foi indicada pela não rejeição do teste de *Breusch-Pagan* ao nível de significância de 5%.

Figura 3.6: Resíduos versus valores preditos e resíduos versus quantis teóricos, para o modelo sem interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019

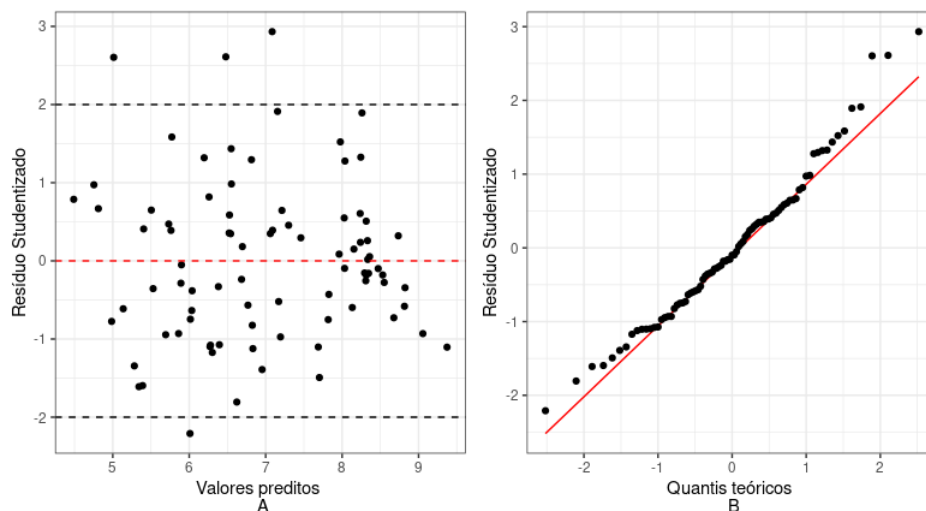
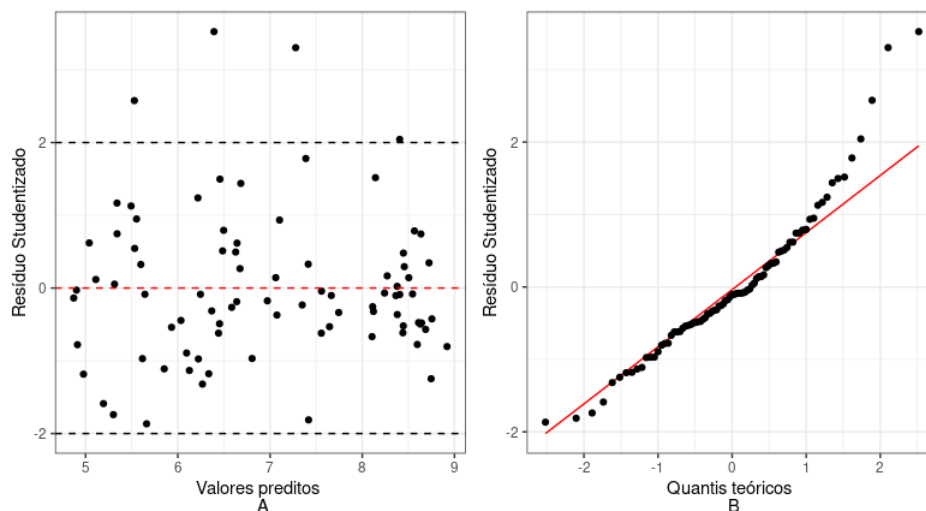


Figura 3.7: Resíduos versus valores preditos e resíduos versus quantis teóricos, para o modelo com interação da variável continente, em que as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019



3.5 Análise de observações não usuais e multicolinearidade

Nas análises que foram feitas para detectar pontos influentes ou outliers (ver Apêndice B), é possível notar que as observações 66 e 72 aparecem em mais gráficos e medidas como pontos *outliers* que impactam demasiadamente as estimativas dos coeficientes, os valores preditos e a matriz de variância e covariância das estimativas. Essas observações

representam os países de Portugal e Sérvia, respectivamente.

Ao observar os valores de VIF na Tabela 3.6, é possível notar que não há nenhum valor de VIF maior que 10, desse modo, é razoável dizer que não há multicolinearidade significativa no modelo de regressão linear selecionado.

O maior valor do coeficiente de determinação dos modelos em a variável explicativa é a variável resposta em função das demais variáveis explicativas foi de 0,59. Portanto, não é o caso que as variáveis explicativas são totalmente não correlacionadas, porém, elas não são correlacionadas ao ponto de causar um problema de multicolinearidade relevante.

Tabela 3.6: Coeficiente de determinação e valores VIF para as variáveis explicativas do modelo de regressão da Tabela 3.2

Variável	R_j^2	VIF
QI	0,57	2,35
Educação	0,49	1,97
ExpectHomens	0,59	2,43
ContinenteÁfrica	0,48	1,92
ContinenteAmérica	0,31	1,45
ContinenteAsia	0,41	1,71

3.6 Interpretação do Modelo

O modelo final escolhido foi o sumarizado na Tabela 3.2 e na Tabela 3.7, pois ele não violou as suposições normalidade, homoscedasticidade e independência dos erros. Ademais, foi escolhido como o melhor modelo pelo teste F parcial via *backward*.

Tabela 3.7: Modelo de regressão, selecionado através do método *backward* baseado no teste F parcial, em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal em vários países em 2019

	Coefficiente	EP	Est. T	Valor-p
Intercepto	-4,768	1,049	-4,544	< 0,001
QI	0,020	0,009	2,394	0,019
Educacao	0,000	0,0001	2,914	0,005
ExpectHomens	0,135	0,015	8,927	< 0,001
Continente:				
África	-0,057	0,267	-0,215	0,831
América	0,252	0,198	1,269	0,208
Asia	-0,337	0,157	-2,150	0,035

Desse modo, a interpretação que as variáveis do modelo tem com a variável renda mensal é:

- Mantendo-se todas as variáveis constantes, o acréscimo de uma unidade no QI é responsável por um aumento de 2,05% na renda mensal.
- Mantendo-se todas as variáveis constantes, o acréscimo de uma unidade na variável Educação é responsável por um aumento de 0,02% na renda mensal.
- Mantendo-se todas as variáveis constantes, o acréscimo de uma unidade na expectativa de vida masculina é responsável por um aumento de 14,40% na renda mensal.
- Quando o país pertence a África, é esperado uma redução de 5,57% no valor da renda mensal quando comparado aos países da Europa.
- Quando o país pertence a América, é esperado um aumento de 28,63% no valor da renda mensal quando comparado aos países da Europa.
- Quando o país pertence a Ásia, é esperado uma redução de 28,64% no valor da renda mensal quando comparado aos países da Europa.

Ao observar a Tabela 3.8, o modelo selecionado indicou que o Japão tem um maior valor de renda mensal do que Alemanha e Estados Unidos, o que não é observado na realidade. Há um erro relevante na estimativa de renda mensal para os Estados Unidos

Tabela 3.8: Valores preditos de renda mensal para diferentes países

País	QI	Educação	ExpectHomens	Renda Estimada	Renda Observada
Brasil	83	427	70,7	869,69	645,00
Japão	106	1341	81,6	4063,58	3554,00
Alemanha	100	1883	78,6	3789,01	4305,00
Camarões	67	36	59,1	89,07	133,00
EUA	97	2608	74,5	3092.92	5911,00

da América. Esses erros, muito provavelmente, é devido ao peso que se foi dado no modelo para expectativa de vida dos homens. Portanto, um país com elevada expectativa de vida masculina, como é o caso do Japão, tenderá a ter uma renda estimada relativamente maior do que um país com baixa expectativa de vida masculina.

4 Conclusão

Neste relatório foi possível relacionar a renda mensal com outras variáveis socioeconômicas, demográficas e climáticas presentes no conjunto de dados considerado. Passando de análise descritiva das variáveis, transformação, seleção de modelos (e variáveis) e análise de resíduos.

Na análise descritiva, foi possível observar que as variáveis despesas com educação e renda mensal apresentaram uma distribuição assimétrica à direita, o que expressa a discrepância já conhecida entre os países – poucos países tem uma renda mensal por habitante e gastos com a educação altos. Quando condicionado pelo continente, foi possível perceber ainda mais discrepâncias tanto nas distribuições das variáveis em cada continente quanto a relação entre as variáveis, evidenciando que o continente é um fator de influência tanto na variável renda mensal quanto nas variáveis explicativas.

Ademais, foi notado a relação exponencial com as variáveis despesas com educação, expectativa de vida dos homens e das mulheres, e natalidade com a variável renda mensal. Dessa forma, foi necessário realizar a transformação Box-Cox na variável resposta, sendo que a transformação logarítmica foi a mais apropriada.

Para selecionar o modelo de regressão linear múltiplo com as variáveis mais importantes para prever a renda mensal, foi avaliado o teste F parcial via *backward*, AIC e o coeficiente de determinação ajustado, sendo os dois últimos realizados para todas as combinações possíveis de modelos com as variáveis explicativas do conjunto de dados.

O modelo escolhido, após o uso do gráfico da regressão parcial, foi avaliado considerando interação da variável categórica continente com as outras variáveis explicativas. Apesar do teste F parcial indicar que o modelo com interação é significativo, esse modelo apresentou problema nos resíduos, uma vez que a suposição de normalidade dos resíduos foi violada. Em contraste, o modelo aditivo, que foi selecionado anteriormente, não apresentou problemas nos resíduos.

Na análise de observações não usuais, foi possível notar que os países de Portugal e Sérvia apresentam-se como observações influentes, já que foram indicadas como influentes pelas métricas de Distância de Cook, DFFITS, COVARATIO. Desse modo, essas países estão impactando de modo demasiado na estimativas de coeficientes, nos valores preditos

e até na matriz de covariância das estimativas.

Ao analisar a possível multicolinearidade através da métrica VIF, foi observado que, para nenhuma variável, o valor de VIF não foi maior que 10, desse modo, não indicando um grave problema de multicolinearidade nos dados. Pelos valores que foram obtidos para os coeficientes de determinação para cada variável explicativa, não é possível dizer que as variáveis estão plenamente não correlacionadas entre si, todavia, também não é possível dizer que há um grave problema de multicolinearidade entre as variáveis.

Além disso, uma discussão mais aprofundada das implicações econômico-sociais que os resultados desse estudo podem fornecer.

As variáveis explicativas do modelo aditivo foram interpretadas à luz da variável resposta renda mensal. Um ano a mais na expectativa de vida do homens está associado com um aumento de 14,40% na renda mensal. Em relação a categoria de referência Europa, estar na América representou um aumento de 28,63% e estar na Ásia apresentou uma redução de 28,64% na renda mensal. O modelo atribuiu uma importância elevada a expectativa de vida masculina, o que foi evidenciado pelo contraste das estimativas da renda mensal em Japão e Estados Unidos com os valores reais.

REFERÊNCIAS

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.
- Biciunaite, A. (2018). Aumento da expectativa de vida depende da queda da linha da pobreza. <https://plenae.com/artigos/aumento-da-expectativa-de-vida-depender-da-queda-da-linha-da-pobreza/>. Acesso em 10 de Julho de 2023.
- Box, G. E. e Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2), 211–243.
- Breusch, T. S. e Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, p. 1287–1294.
- Chancel, L., Piketty, T., Saez, E. e Zucman, G. (2022). *World inequality report 2022*. Harvard University Press.
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15–18.
- Demétrio, C. G. B. e Zocchi, S. S. (2006). Modelos de regressão. *Piracicaba: ESALQ*.
- Durbin, J. e Watson, G. S. (1992). *Testing for serial correlation in least squares regression. I*. Springer.
- Fox, J. (1991). *Regression diagnostics: an introduction (Quantitative Applications in the Social Sciences)*. Sage publications.
- Gujarati, D. N. e Porter, D. C. (2011). *Econometria básica*. Amgh Editora.
- Hoffmann, R. (2016). *Análise de regressão: uma introdução à econometria*. Portal de Livros Abertos da USP.
- Nargund, G. (2009). Declining birth rate in developed countries: A radical policy re-think is required. *Facts, views & vision in ObGyn*, **1**(3), 191.

REFERÊNCIAS

- Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. *et al.* (1996). *Applied linear statistical models*. Irwin Chicago.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Shapiro, S. S. e Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3/4), 591–611.
- Worlddata (2019). Worlddata: The world in numbers. <https://www.worlddata.info/>. Acesso em 17 de Abril de 2023.

A Gráfico de dispersão das variáveis de acordo com o continente

Figura A.1: Gráfico de dispersão, densidade e correlação das variáveis socioeconômicas, demográficas e climáticas dos países em 2019 de acordo com o continente



B Gráficos para identificação de observações não usuais

Figura B.1: Gráfico da distância de *cook* e resíduos estudentizados versus *leverage* da regressão linear em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para predizer a renda mensal em vários países em 2019

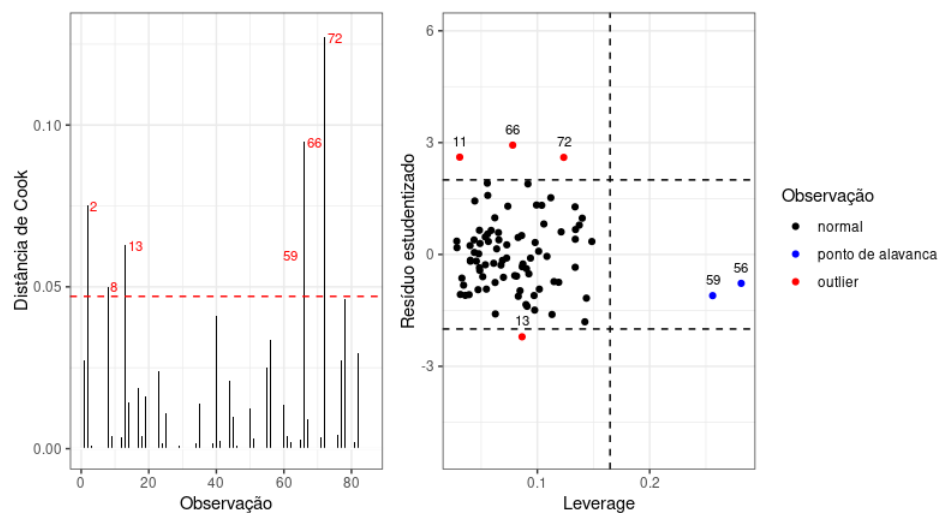


Figura B.2: Gráfico dos DFFITS e COVRATIO da regressão linear em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019

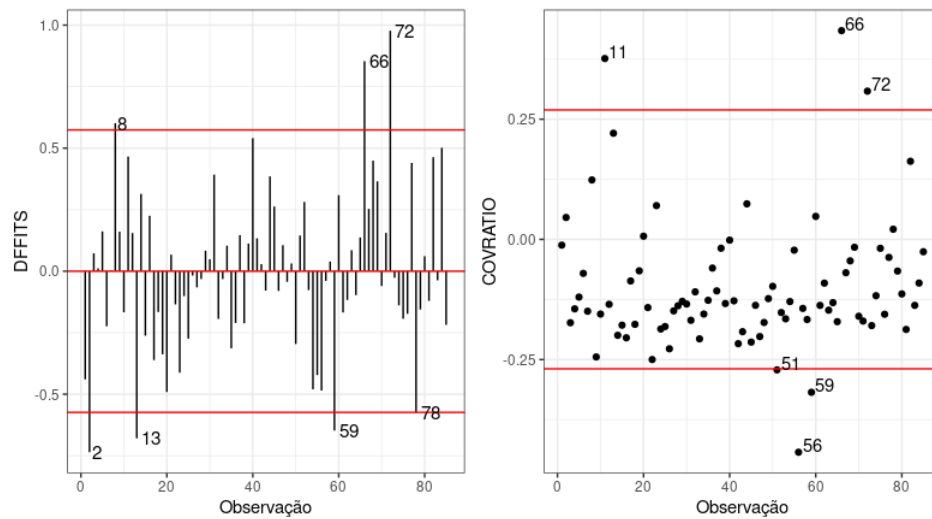
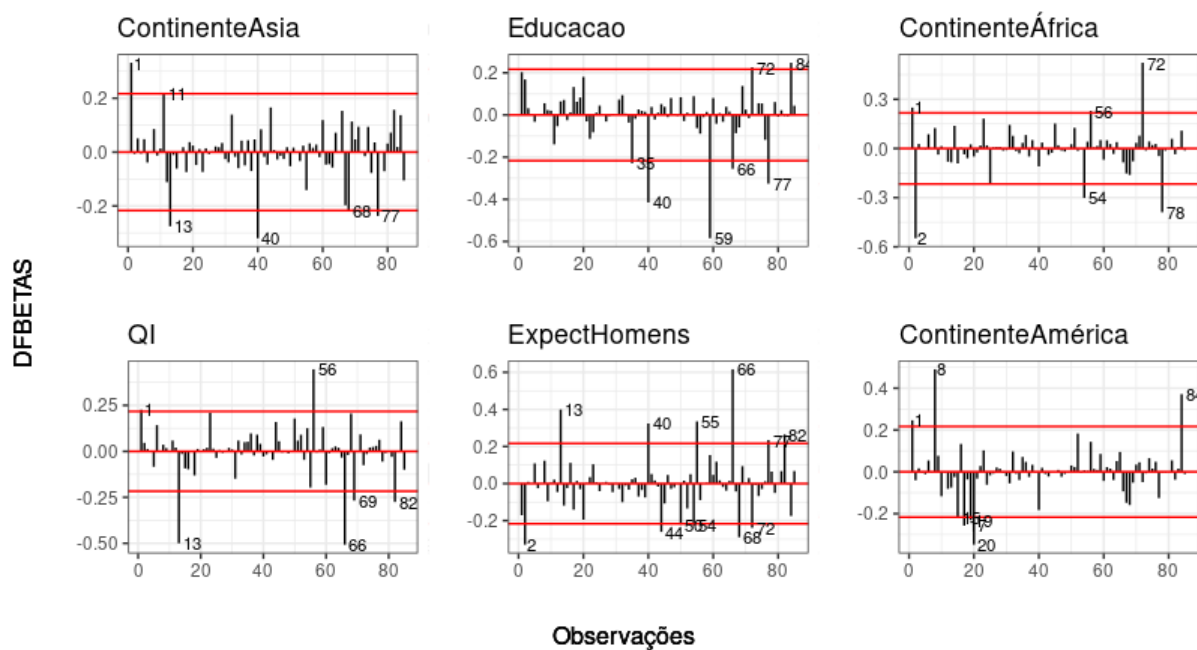


Figura B.3: Gráfico dos DFBETAS da regressão linear em as variáveis socioeconômicas, demográficas e climáticas são utilizadas para prever a renda mensal em vários países em 2019



C Códigos

```
library(tidyverse)
library(GGally)
library(corrplot)
library(patchwork)
library(olsrr)
library(gridExtra)
library(gmodels)
library(lindia)
library(qqplotr)
library(dplyr)
library(lmtest)
library(MASS)
library(maps)
```

```
df <- read.csv("dados_mundiais.csv")[c(-54,-60),-1]
```

```
#mapa
# paises <-
#
# mapa_mundo <- map_data("world")
# mapa_df <- data.frame(mapa_mundo, presente = ifelse(mapa_mundo$region %in% paises
#
# ggplot(data = mapa_df, aes(x = long, y = lat, group = group)) +
#   geom_polygon(aes(fill = presente)) +
#   scale_fill_manual(values = c("darkblue", "lightblue"), guide = FALSE) +
#   coord_quickmap() +
#   theme_void()
```

```
#Analise descritiva
```

```
df <- df|>
```

```

mutate(Continente = as.factor(Continente))

df$Continente = relevel(df$Continente, "Europa")

df1 <- subset(df, df$Continente != "Oceania")

df1<- df1%>%relocate(RendaMensal, .before=Continente)

sqrt(diag(var(df[2:9]))) / colMeans(df[2:9], na.rm = TRUE)
summary(df[2:9])
var(df[2:9], na.rm = TRUE)
var(df[2:9])

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

apply(df[2:9], 2, Mode)

ggpairs(df1[, 2:(length(df)-1)],
  aes(color=df1$Continente),
  lower = list(continuous = "blank") ,
  diag = list(continuous = "blankDiag"))

ggpairs(df1[, 2:(length(df)-1)])

ggpairs(df[, 2:(length(df)-1)], lower = list(continuous = "blank"),
  upper = list(continuous = "points") ,
  diag = list(continuous = "blankDiag"))

df1<- df
df1<- df1 %>%
  relocate(RendaMensal, .before=Continente)

```

```
df1[c(3,58),'Continente'] = 'Asia'
```

```
boxp <- function(y,name){
  ggplot(df)+
    geom_boxplot(aes(y=df[,y+1],x=Continente,color=Continente))+
    labs(y=name)+
    theme(legend.position="none")+
    scale_color_manual("Continente",breaks=NULL,
                      values=c("#ff6961","#77dd77","#e69c0d",
                                "#84b6f4","#cd9cb2"))+
    scale_x_discrete("",breaks = NULL) +
    theme_bw()
}
```

```
boxp1 <- function(y,name){
  ggplot(df)+
    geom_boxplot(aes(y=df[,y+1],x=Continente,color=Continente))+
    labs(y=name)+
    theme(legend.position="none")+
    scale_color_manual("Continente",
                      values=c("#ff6961","#77dd77","#e69c0d",
                                "#84b6f4","#cd9cb2"))+
    scale_x_discrete("",breaks = NULL) +
    theme_bw()
}
```

```
boxp(2,"QI") / boxp(3,"Despesas com educação") |
  boxp(4,"Temperatura máxima diária") /
  boxp(1,"Renda mensal") | boxp(5,"Expectativa de vida masculina") /
  boxp(6,"Expectativa de vida feminina") | boxp1(7,"Taxa de natalidade") /
  boxp1(8,"Taxa de mortalidade")
```

```
# Transformações
```

```
fit_orig <- lm(RendaMensal~., data = df1[,-1])
bc <- boxcox(RendaMensal~.,data=df1[,-1],ylab='Log-verossimilhança')
plot(bc)
bc$x[which.max(bc$y)]
```

```

# Modelo
fit <- lm(log(RendaMensal)~., data = df1[,-1])
summary(fit)
shapiro.test(fit$residuals)
gqtest(fit)
dwtest(fit)

MyAnova <- function(modelo, nd){
  m1 <- modelo
  np <- dim(anova(m1))[1]
  SQReg <- round(sum(anova(m1)$"Sum Sq"[1:(np-1)]), nd)
  glReg <- sum(anova(m1)$"Df"[1:(np-1)])
  SQRes <- round(anova(m1)$"Sum Sq"[np], nd)
  glRes <- anova(m1)$"Df"[np]
  SQTotal <- round(SQReg + SQRes, nd)
  glTotal <- glReg + glRes
  QMReg <- round(SQReg/glReg, nd)
  QMRes <- round(SQRes/glRes, nd)
  MyF <- round(QMReg/QMRes, nd)
  vpF <- ifelse(pf(MyF, glReg, glRes, lower.tail = F) < 0.0001, "<0.001",
               roud(pf(MyF, glReg, glRes, lower.tail = F), nd))
  ncolunas <- c("Fonte de Variação", "SQ", "gl", "F", "valor p")
  Tanova <- data.frame(FV = c("Regressão",
                              "Resíduos",
                              "Total"),
                      gl = c(glReg, glRes, glTotal),
                      SQ = c(SQReg, SQRes, SQTotal),
                      QM = c(QMReg, QMRes, " "),
                      Est.F = c(MyF, " ", " "),
                      valor.p = c(vpF, " ", " ")
  )
  Tanova
}

MyAnova(fit,2)

# selecao de modelos
summary(step(fit))

```

```

summary(step(fit, k = log(nrow(df))))

ols_step_best_subset(fit)
ols_step_backward_aic(fit)

all <- data.frame(ols_step_all_possible(fit))
arrange(all, desc(adjr))
head(arrange(all, desc(adjr)))

# Teste F parcial
anova(fit, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens + ExpectMulheres +
               Natalidade + Mortalidade, data=df1[, -1]))
anova(fit, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens + ExpectMulheres +
               Natalidade + Continente, data=df1[, -1]))
anova(fit, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens + ExpectMulheres +
               Mortalidade + Continente, data=df1[, -1]))
anova(fit, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens +
               Natalidade + Mortalidade + Continente, data=df1[, -1]))
anova(fit, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectMulheres +
               Natalidade + Mortalidade + Continente, data=df1[, -1]))
anova(fit, lm(log(RendaMensal) ~ QI + Educacao + ExpectHomens + ExpectMulheres +
               Natalidade + Mortalidade + Continente, data=df1[, -1]))
anova(fit, lm(log(RendaMensal) ~ QI + TempMax + ExpectHomens + ExpectMulheres +
               Natalidade + Mortalidade + Continente, data=df1[, -1]))
anova(fit, lm(log(RendaMensal) ~ Educacao + TempMax + ExpectHomens + ExpectMulheres +
               Natalidade + Mortalidade + Continente, data=df1[, -1]))

# sai expect feminina
fit2 <- lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens +
            Natalidade + Mortalidade + Continente, data=df[, -1])

anova(fit2, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens +
                Natalidade + Mortalidade, data=df1[, -1]))
anova(fit2, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens +
                Natalidade + Continente, data=df1[, -1]))
anova(fit2, lm(log(RendaMensal) ~ QI + Educacao + TempMax + ExpectHomens +
                Mortalidade + Continente, data=df1[, -1]))
anova(fit2, lm(log(RendaMensal) ~ QI + Educacao + TempMax +
                Natalidade + Mortalidade + Continente, data=df1[, -1]))

```

```

anova(fit2,lm(log(RendaMensal)~QI+Educacao+ExpectHomens+
               Natalidade+Mortalidade+Continente, data=df1[,,-1]))
anova(fit2,lm(log(RendaMensal)~QI+TempMax+ExpectHomens+
               Natalidade+Mortalidade+Continente, data=df1[,,-1]))
anova(fit2,lm(log(RendaMensal)~Educacao+TempMax+ExpectHomens+
               Natalidade+Mortalidade+Continente, data=df1[,,-1]))
# sai temp max
fit3 <- lm(log(RendaMensal)~QI+Educacao+ExpectHomens+ Natalidade
           +Mortalidade+Continente, data=df[,,-1])

anova(fit3,lm(log(RendaMensal)~QI+Educacao+ExpectHomens+
               Natalidade+Mortalidade, data=df1[,,-1]))
anova(fit3,lm(log(RendaMensal)~QI+Educacao+ExpectHomens+
               Natalidade+Continente, data=df1[,,-1]))
anova(fit3,lm(log(RendaMensal)~QI+Educacao+ExpectHomens+
               Mortalidade+Continente, data=df1[,,-1]))
anova(fit3,lm(log(RendaMensal)~QI+Educacao+
               Natalidade+Mortalidade+Continente, data=df1[,,-1]))
anova(fit3,lm(log(RendaMensal)~QI+ExpectHomens+
               Natalidade+Mortalidade+Continente, data=df1[,,-1]))
anova(fit3,lm(log(RendaMensal)~Educacao+ExpectHomens+
               Natalidade+Mortalidade+Continente, data=df1[,,-1]))

# sai mortalidade
fit4 <- lm(log(RendaMensal)~QI+Educacao+ExpectHomens+ Natalidade+Continente,
           data=df[,,-1])

anova(fit4,lm(log(RendaMensal)~QI+Educacao+ExpectHomens+
               Natalidade, data=df1[,,-1]))
anova(fit4,lm(log(RendaMensal)~QI+Educacao+ExpectHomens+Continente,
               data=df1[,,-1]))
anova(fit4,lm(log(RendaMensal)~QI+Educacao+
               Natalidade+Continente, data=df1[,,-1]))
anova(fit4,lm(log(RendaMensal)~QI+ExpectHomens+
               Natalidade+Continente, data=df1[,,-1]))
anova(fit4,lm(log(RendaMensal)~Educacao+ExpectHomens+
               Natalidade+Continente, data=df1[,,-1]))

#Sai Natalidade

```

```

fit5 <- lm(log(RendaMensal)~QI+Educacao+ExpectHomens+Continente, data=df1[,-1])

anova(fit5,lm(log(RendaMensal)~QI+Educacao+ExpectHomens, data=df1[,-1]))
anova(fit5,lm(log(RendaMensal)~QI+Educacao+Continente, data=df1[,-1]))
anova(fit5,lm(log(RendaMensal)~QI+ExpectHomens+Continente, data=df1[,-1]))
anova(fit5,lm(log(RendaMensal)~Educacao+ExpectHomens+Continente, data=df1[,-1]))

summary(fit5)

fit_51 <- lm(Natalidade~QI+Educacao+ExpectHomens+Continente, data=df1[,-1])

ggplot()+
  aes(x = fit_51$residuals, y = fit5$residuals) +
  geom_point()+
  geom_smooth(method = lm,se = FALSE) +
  labs(x = "Resíduos Natalidade|. ",y="Resíduos log(Renda)|. ")+
  theme_bw()

summary(lm(fit5$residuals~fit_51$residuals))

# Teste de interação no modelo
fit6 <- lm(log(RendaMensal)~(QI+Educacao+ExpectHomens)*Continente, data=df1[,-1])
summary(fit6)
anova(lm(log(RendaMensal)~QI+Educacao+ExpectHomens+Continente,data=df1[,-1]),fit6)

# Modelo selecionado
model1 <- lm(log(RendaMensal)~QI+Educacao+ExpectHomens+Continente,data=df1[,-1])
model2<- fit6

# Analise dos residuos
shapiro.test(model1$residuals)
gqtest(model1)
bptest(model1)
dwtest(model1)

shapiro.test(model2$residuals)
gqtest(model2)

```



```

bptest(model2)
dwtest(model2)
b <- ggplot(model1)+
  geom_point(aes(x=.fitted,y=studres(model1)))+
  geom_hline(yintercept = c(2,-2), colour = "black",linetype = "dashed")+
  geom_hline(yintercept = c(0), colour = "red",linetype = "dashed")+
  scale_color_discrete("Continente")+
  labs(x="Valores preditos \n A",y="Resíduo Studentizado")+
  theme_bw()

```

```

a <- ggplot(model1)+
  aes(sample=studres(model1))+
  geom_qq_line(color="red")+
  #geom_qq_band(fill="lightblue")+
  geom_qq()+
  labs(x="Quantis teóricos \n B", y="Resíduo Studentizado")+
  theme_bw()

```

b | a

```

b <- ggplot(model2)+
  geom_point(aes(x=.fitted,y=studres(model2)))+
  geom_hline(yintercept = c(2,-2), colour = "black",linetype = "dashed")+
  geom_hline(yintercept = c(0), colour = "red",linetype = "dashed")+
  scale_color_discrete("Continente")+
  labs(x="Valores preditos \n A",y="Resíduo Studentizado")+
  theme_bw()

```

```

a <- ggplot(model2)+
  aes(sample=studres(model2))+
  geom_qq_line(color="red")+
  #geom_qq_band(fill="lightblue")+
  geom_qq()+
  labs(x="Quantis teóricos \n B", y="Resíduo Studentizado")+
  theme_bw()

```

b | a

```
# Análise de Multicolinearidade
```

```
model <- model1
```

```
1-1/vif(model1)
```

```
# Analise de observações nao usuais
```

```
model <- model1
```

```
#leverage
```

```
g <- ols_prep_rstudlev_data(model)
```

```
d <- g$levrstud
```

```
d$txt <- ifelse(d$color == "normal", NA, d$obs)
```

```
f <- d[d$color == "outlier", c("obs", "leverage", "rstudent")]
```

```
colnames(f) <- c("observation", "leverage", "stud_resid")
```

```
a <- ggplot(d, aes(leverage, rstudent, label = txt)) +
  geom_point(aes(colour = fct_color)) +
  scale_color_manual(labels = c("normal", "ponto de alavanca", "outlier",
                                "outlier e ponto de alavanca"),
                     values = c("black", "blue", "red", "green")) +
  xlim(g$minx, g$maxx) +
  ylim(g$miny, g$maxy) +
  labs(colour = "Observação", x = "Leverage", y = "Resíduo estudentizado") +
  geom_hline(yintercept = c(2,-2), colour = "black", linetype = "dashed") +
  geom_vline(xintercept = g$lev_thrsh, colour = "black", linetype = "dashed") +
  geom_text(vjust = -1, size = 3, colour = "black") +
  theme_bw()
```

```
# Distancia de Cook
```

```
k <- ols_prep_cdplot_data(model)
```

```
d <- ols_prep_outlier_obs(k)
```

```
f <- ols_prep_cdplot_outliers(k)
```

```
b <- ggplot(d, aes(x = obs, y = cd, label = txt)) +
  geom_bar(width = 0.2, stat = "identity",
           aes(fill = fct_color)) +
  scale_fill_manual(values = c("black", "black"), breaks=NULL) +
  ylim(0, k$maxx) +
  labs(x = "Observação", y = "Distância de Cook") +
```

```

geom_hline(yintercept = 0, colour = "white") +
geom_hline(yintercept = k$ts, colour = "red", linetype = "dashed") +
geom_text(hjust = -0.2, nudge_x = 0.05, size = 3, na.rm = TRUE, color="red") +
theme_bw()

```

b | a

```

ols_dfbetas <- function (model, print_plot = TRUE)
{
  obs <- NULL
  txt <- NULL
  dfb <- dfbetas(model)
  n <- nrow(dfb)
  np <- ncol(dfb)
  threshold <- 2/sqrt(n)
  myplots <- list()
  outliers <- list()
  for (i in 2:seqs) {
    dbetas <- dfb[, i]
    df_data <- data.frame(obs = seq_len(n), dbetas = dbetas)
    d <- ols_prep_dfbeta_data(df_data, threshold)
    f <- ols_prep_dfbeta_outliers(d)
    p <- ggplot(d, aes(x = obs, y = dbetas,
                      label = txt, ymin = 0, ymax = dbetas)) +
      geom_linerange(colour = "black") +
      geom_hline(yintercept = c(0, threshold, -threshold),
                 colour = "red") +
      labs(x="", y="") +
      ggtitle(paste(colnames(dfb)[i])) +
      theme_bw()+
      geom_text(hjust = -0.2, nudge_x = 0.15, size = 3,
                colour = "black", na.rm = TRUE)
    myplots[[i]] <- p
    outliers[[i]] <- f
  }
  if (print_plot) {
    marrangeGrob(myplots, nrow = 2, ncol = 3, top = quote(paste("")),
                  left="DFBETAS",bottom="Observações")
  }
}

```

```
}
```

```
ols_dfbetas(model)
```

```
#DFFITS
```

```
dbetas <- NULL
```

```
obs <- NULL
```

```
txt <- NULL
```

```
dffitsm <- unlist(dffits(model))
```

```
k <- length(coef(model))
```

```
n <- nrow(df)
```

```
dffits_t <- sqrt(k/n) * 2
```

```
title <- names(model.frame(model))[1]
```

```
dfits_data <- data.frame(obs = seq_len(n), dbetas = dffitsm)
```

```
d <- ols_prep_dfbeta_data(dfits_data, dffits_t)
```

```
f <- ols_prep_dfbeta_outliers(d)
```

```
a <- ggplot(d, aes(x = obs, y = dbetas, label = txt, ymin = 0,
                  ymax = dffitsm)) +
  geom_linerange(colour = "black") +
  geom_hline(yintercept = c(0, dffits_t, -dffits_t), colour = "red") +
  labs(x = "Observação", y = "DFFITS") +
  theme_bw()+
  geom_text(hjust = -0.2, nudge_x = 0.15, size = 4, colour = "black",
            na.rm = TRUE)
```

```
#COVRATIOs
```

```
cr = data.frame(covratio(model))
```

```
cr['obs'] = seq(1,nrow(df))
```

```
limite = (3*length(model$coefficients) ) / (nrow(df)-length(model$coefficients))
```

```
b <- ggplot(data=cr,aes(x = obs, y = (1-covratio.model.), label = obs))+
  geom_point() +
  geom_hline(yintercept = c(limite, -1*limite),colour = "red")+
  geom_text(aes(label=ifelse(abs(1-covratio.model.)>limite,as.character(obs),'')),
            hjust=-0.2,vjust=0) +
  labs(x="Observação",y="COVRATIO")+
  theme_bw()
```

```
a | b
```