
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 parcial 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

# LISTA DE TABELAS

# SUMÁRIO

# 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, aumentar a criminalidade e a instabilidade política. Além disso, a desigualdade de renda pode levar a disparidades na saúde e no acesso a outros serviços básicos, como educação e moradia. Ela também pode limitar as oportunidades de desenvolvimento econômico e reduzir o crescimento sustentável a longo prazo.

A redução da desigualdade de renda 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.

Diante disso, 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.
- Quanificar 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 natalidade, taxa de mortalidade, 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 e o acesso a cuidados de saúde. Países com renda mais alta geralmente podem ter uma expectativa de vida maior devido ao acesso a melhores cuidados de saúde, alimentação adequada e condições de vida melhores. A inclusão dessa variável permite investigar como a longevidade está relacionada à renda mensal.

Em relação à taxa de natalidade, em países com renda mais alta, as taxas tendem a ser mais baixas devido ao acesso à educação, métodos contraceptivos e oportunidades de carreira. A taxa de natalidade influencia a composição demográfica e pode impactar a renda média da população.

Além disso, 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. A taxa de mortalidade é um indicador importante para avaliar a qualidade de vida e os desafios de saúde enfrentados por diferentes grupos populacionais.

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 sistêmicas e estruturais 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.

A contextualização de vocês é excelente. Mas, entendo que outras referências poderiam ser incluídas, principalmente sobre as justificativas de inclusão de variáveis na análise.

# 2 Materiais e métodos

Os métodos descritos nessa seção serão utilizados nos dados de diversos países, todas as análises serão realizadas com o auxílio do *Software R* (R Core Team, 2022). Primeiramente, será realizada uma análise da relação entre a renda mensal e as variáveis explicativas. Caso essa relação não seja linear, uma transformação da renda mensal será proposta com base na transformação Box-Cox.

Em seguida, modelos de regressão linear múltiplo serão utilizados para modelar a relação da renda mensal com as variáveis explicativas selecionadas através do método de seleção de variáveis *backward* e todas as regressões possíveis baseado no AIC e no teste F parcial. A significância global e individual dos parâmetros serão observadas, bem como a proporção da variabilidade original dos dados explicada pelos modelos. E por fim, serão analisados os resíduos dos modelos para avaliar sua qualidade e verificar se os pressupostos da regressão linear estão sendo satisfeitos.

## 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.

## 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 $k$ variáveis explicativas é:

$$\boldsymbol{Y}_{(n\text{x}1)} = \boldsymbol{X}_{(n\text{x}p)}\boldsymbol{\beta}_{(p\text{x}1)} + \boldsymbol{\epsilon}_{(n\text{x}1)} \tag{2.1}$$

em que $p = k+1$; $\boldsymbol{Y}$ é o vetor da variável resposta; $\boldsymbol{X}$ contém as observações das variáveis explicativas, bem como uma coluna adicional de 1's, para representar o coeficiente $\beta_0$; $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos; $\epsilon$ é o vetor de variáveis aleatórias não observáveis (erros).

Dessa forma, têm-se as suposições:

i. a variável resposta é função linear das variáveis explicativas;

ii. as variáveis explicativas são fixas;

iii. $\text{E}(\epsilon_i) = 0$, ou seja, $\text{E}(\epsilon) = \mathbf{0}$, sendo $\mathbf{0}$ um vetor de zeros de dimensões $n$x1;

iv. os erros são homocedásticos, $\text{Var}(\epsilon_i) = \text{E}(\epsilon_i^2) = \sigma^2$;

v. os erros são independentes, $\text{Cov}(\epsilon_i,\epsilon_i') = \text{E}(\epsilon_i,\epsilon_i') = 0$, $i \neq i'$;

vi. 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 $\beta$.

O método consiste em estimar os parâmetros $\beta_j$ de tal forma que os desvios dos valores observados em relação aos estimados sejam mínimos, isso equivale a minimizar o comprimento do vetor $\epsilon$. Usando a norma euclidiana para avaliar ~~com~~ comprimento desse vetor, tem-se:

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

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

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

Assim, para compreender o impacto das variáveis explicativas na variável resposta é possível interpretar ~~d~~os 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 hipótese para os parâmetros

Para testar a significância conjunta dos parâmetros do modelo, isto é, testar $H_0 : \beta_1 = \ldots = \beta_k = 0$ versus $H_1$ : pelo menos um $\beta_j$ 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

$$SQTotal = \boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y}$$

$$SQReg = \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y}$$

$$SQRes = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$$

em que $\boldsymbol{J}$ é uma matriz $n$ x $n$ 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 Anova 2.2 mostra esses resultados de análise de variância, bem como os quadrados médios.

Sob a hipótese nula, a estatística F segue uma distribuição F de Snedecor com $k-1$ e

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 | $n-1$ | SQT | | |

$n-k$ 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 $\beta_j$ no modelo, utiliza-se a estatística de teste dada por

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

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

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 concluímos que o coeficiente $\beta_j$ é significativo. Caso contrário, não rejeitamos a hipótese nula e concluímos 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:

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

em que $0 \geq R^2 \geq 1$. 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 $Y_i = \hat{Y}_i \ \forall i$.

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, denotada por $R_a^2$, ele ajusta $R^2$ dividindo cada soma de quadrados por seus graus de liberdade associado:

$$R_a^2 = 1 - \frac{SQRes/(n-k)}{SQTotal/(n-1)} = 1 - \left(\frac{n-1}{n-k}\frac{SQres}{SQTotal}\right)$$

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 *Shapiro-Wilk*.

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

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

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: $X_1 \leq X_2 \leq \cdots \leq X_n$.

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 estatísticos, como o teste de *Bartlett*.

O teste de *Bartlett* (Bartlett, 1937), é um teste estatístico utilizado para avaliar se as variâncias de diferentes grupos de dados são iguais. A estatística de teste $X^2$ do teste de *Bartlett* segue uma distribuição qui-quadrado com $a-1$ graus de liberdade, e sua estatística de teste é dada por:

$$\chi^2 = \frac{(n-a)\ln(S_p^2) - \sum_{i=1}^{a}(n_i-1)\ln(S_i^2)}{1+\frac{1}{3(a-1)}\left(\sum_{i=1}^{a}(\frac{1}{n_i-1}) - \frac{1}{n-a}\right)}$$

onde $S_p^2 = \frac{1}{n-a}\sum_i (n_i-1)S_i^2$ é a estimativa combinada para a variância.

Se o valor de $\chi^2$ for maior que o valor crítico da distribuição qui-quadrado para um determinado nível de significância, a hipótese nula de que as variâncias são iguais é rejeitada.

### 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 ~~pode~~ 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

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

sendo $\lambda$ o parâmetro da transformação e $Y$ a variável resposta, na ausência de uma transformação, $\lambda = 1$. O lambda 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 $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{q*} = 0$, $q* < k$, a estatística de teste é dada por:

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

em que $q$ é o número de parâmetros do modelo restrito -1 e $p = K + 1$ é o número de parâmetros estimados no modelo. Sob a hipótese nula, essa estatística segue uma distribuição F com $p - q - 1$ e $n - p$ 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

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

em que $p$ 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, $b = 1, 2, \ldots, k$ 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.

# 3 Resultados

## 3.1 Análise Descritiva

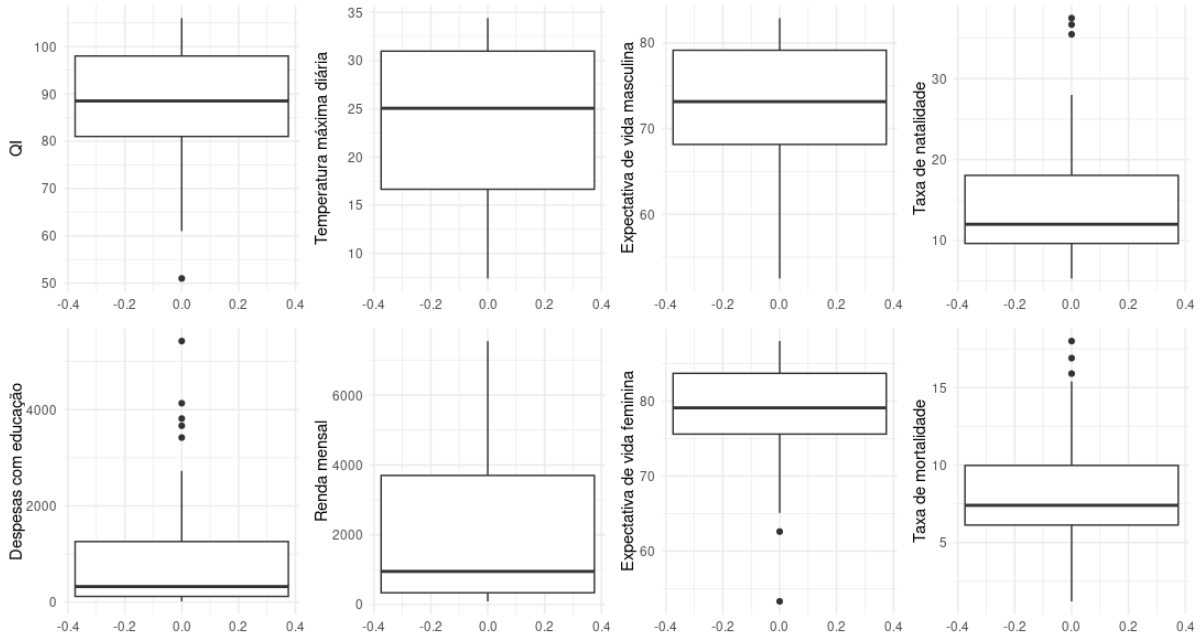
Como pode ser visto na Tabela 3.1, os dados não parecem apresentar simetria se avaliarmos a distância entre os valores de média, moda e mediana. As variáveis da renda mensal e das despesas com educação têm uma alta variabilidade, o que pode ser resultado da diferença de renda e dos gastos com a educação 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 os *boxplots* da Figura 3.1, 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 3.1: *Boxplots* das variáveis socioeconômicas, demográficas e climáticas dos países em 2019



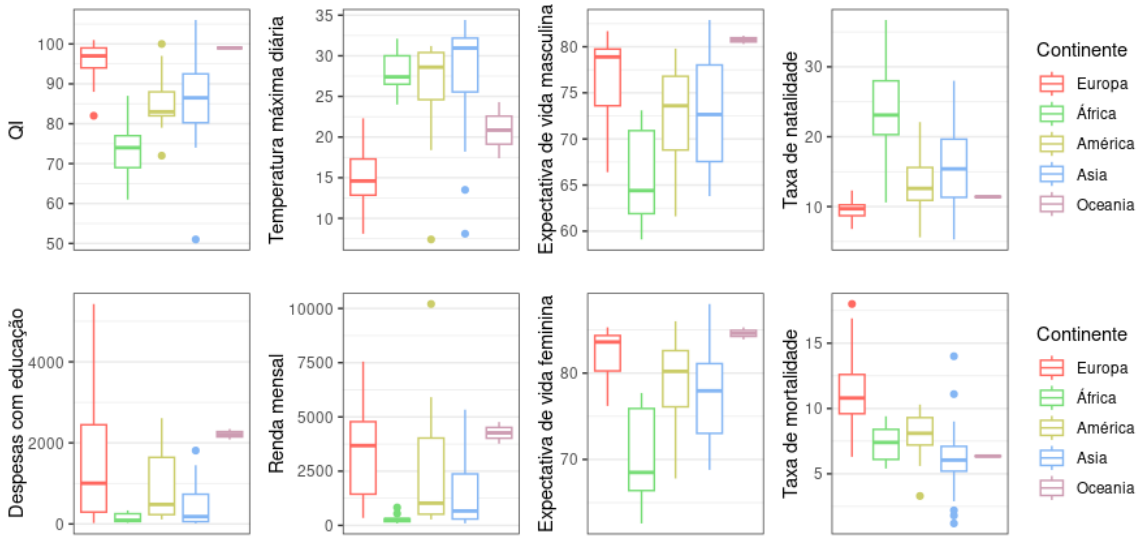

A Oceania tem apenas dois países: Austrália e Nova Zelândia, portanto é mais complicado analisar as características de dispersão das variáveis em relação a esse continente. Mas, ~~o~~ é possível observar ~~no gráfico~~ 3.2~~,~~ é que em relação ao QI, a Oceania apresenta a maior mediana, seguida por 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 3.2: *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.

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 ~~é~~ 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 resposto, isto é algo positivo para o modelo de regressão linear. Mas, 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,

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


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.

Com os gráficos da densidade de cada variável de acordo com o continente, também na Figura 3.3, é possível notar algumas particularidades. A densidade do QI, expectativa de vida masculina e feminina, apresentam uma forte assimetria à esquerda na Europa. Enquanto que a temperatura máxima apresenta uma assimetria à direita também na Europa.

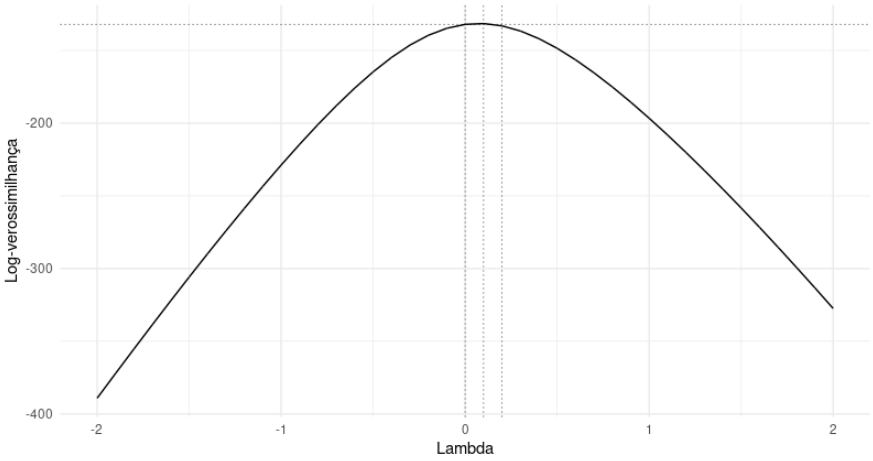
## 3.2 Transformação

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.

Figura 3.4: Log-verossimilhanças para parâmetro da transformação de potência Box-Cox




## 3.3 Seleção de modelos

Como citado na metodologia, será utilizado o teste F parcial, o AIC e $R^2$ ajustado para selecionar as variáveis mais relevantes para predizer 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 $R^2$ foi o modelo da Tabela 3.3.

Tabela 3.2: Modelo de regressão selecionado através do método *backward* baseado no teste F parcial

| | 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 | 0,000 | 0,0001 | 2,914 | **0,005** |
| ExpectHomens | 0,135 | 0,015 | 8,927 | **<0,001** |
| ContinenteÁfrica | -0,057 | 0,267 | -0,215 | 0,831 |
| ContinenteAmérica | 0,252 | 0,198 | 1,269 | 0,208 |
| ContinenteAsia | -0,337 | 0,157 | -2,150 | **0,035** |

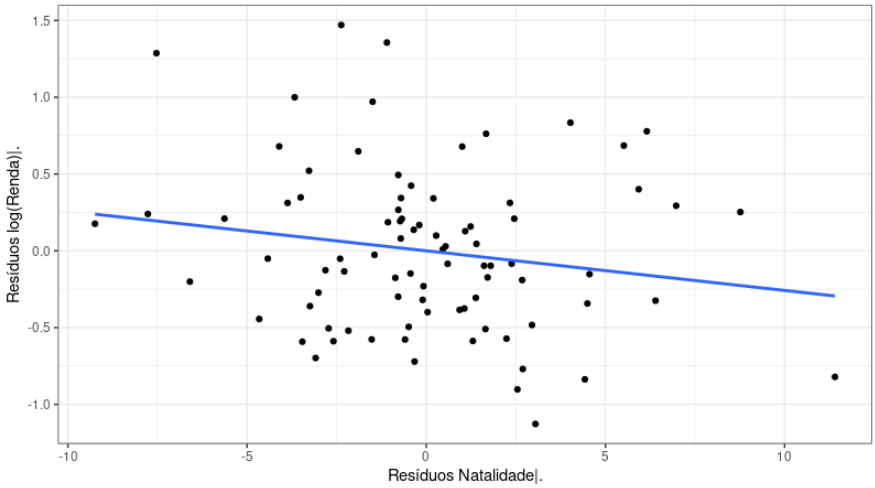
Tabela 3.3: Modelo de regressão selecionado realizando todas as combinações possíveis baseado no AIC e $R^2$ ajustado

| | 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 |
| ContinenteÁfrica | 0,133 | 0,292 | 0,456 | 0,6498 |
| ContinenteAmérica | 0,266 | 0,198 | 1,351 | 0,1808 |
| ContinenteAsia | -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 da 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 para a variável natalidade



### 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 predizer 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 com todas as outras variáveis explicativas

| | 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** |
| ContinenteÁfrica | 12,310 | 3,390 | 3,631 | **<0,001** |
| ContinenteAmérica | 10,050 | 3,681 | 2,731 | **0,008** |
| ContinenteAsia | 2,457 | 2,970 | 0,827 | 0,411 |
| QI:ContinenteÁfrica | -0,069 | 0,038 | -1,821 | 0,073 |
| QI:ContinenteAmérica | -0,089 | 0,037 | -2,405 | **0,019** |
| QI:ContinenteAsia | -0,072 | 0,026 | -2,749 | **0,008** |
| Educacao:ContinenteÁfrica | 0,005 | 0,006 | 3,054 | **0,003** |
| Educacao:ContinenteAmérica | 0,001 | 0,0003 | 3,526 | **0,001** |
| Educacao:ContinenteAsia | 0,0003 | 0,0002 | 1,559 | 0,124 |
| ExpectHomens:ContinenteÁfrica | -0,109 | 0,048 | -2,287 | **0,025** |
| ExpectHomens:ContinenteAmérica | -0,037 | 0,041 | -0,903 | 0,370 |
| ExpectHomens: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*, *Goldfeld-Quandt* 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: Tabela dos testes para as suposições de normalidade, homoscedasticidade e independência dos resíduos para os modelo sem e com interação

| Modelo | Shapiro-Wilk | Goldfeld-Quandt | Durbin-Watson |
|---|---|---|---|
| Sem interação | 0,165 | 0,428 | 0,238 |
| Com interação | **<0,001** | 0,429 | 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 *Goldfeld-Quandt* ao nível de significância de 5%.

Figura 3.6: Resíduos versus valores preditos e resíduos versus quantis teóricos para o modelo sem interação

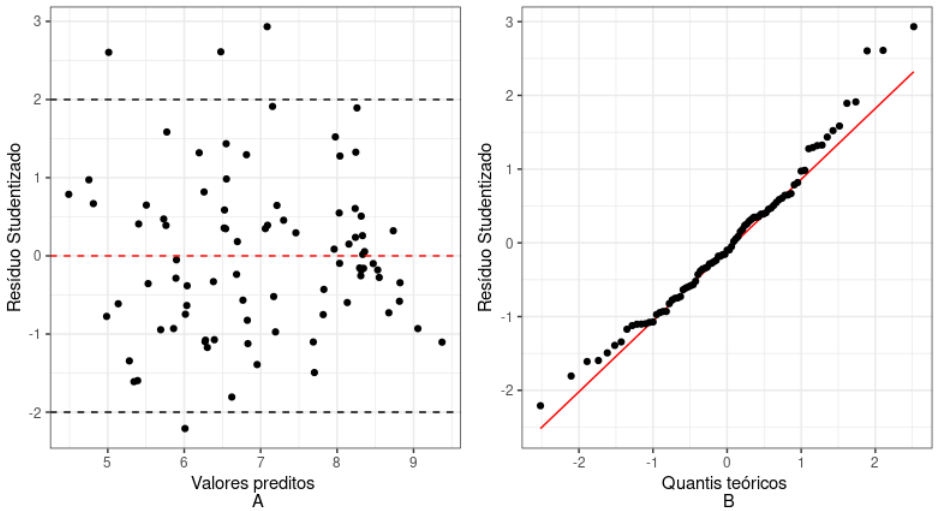
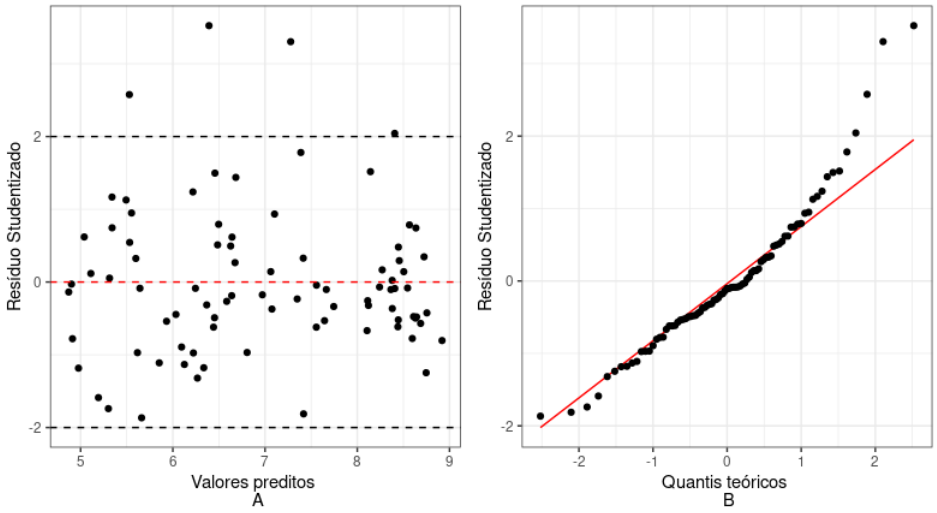


Figura 3.7: Resíduos versus valores preditos e resíduos versus quantis teóricos para o modelo com interação




## 3.5 Interpretação do Modelo

O modelo final escolhido foi o sumarizado na Tabela 3.2 e na Tabela 3.6, 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 3.6: Modelo de regressão selecionado através do método *backward* baseado no teste F parcial

| | 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 | 0,000 | 0,0001 | 2,914 | **0,005** |
| ExpectHomens | 0,135 | 0,015 | 8,927 | **<0,001** |
| ContinenteÁfrica | -0,057 | 0,267 | -0,215 | 0,831 |
| ContinenteAmérica | 0,252 | 0,198 | 1,269 | 0,208 |
| ContinenteAsia | -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.

# 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 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 predizer a renda mensal, foi avaliado p teste F parcial via *backward*, AIC e o $R^2$ 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.

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.

Portanto, o que pode ser feito para o relatório final, é a analisar a multicolinearidade do modelo, já que as variáveis explicativas estão relativamente bem correlacionadas, e realizar um diagnóstico das observações influentes no modelo, ou seja, observações estão afetando bastante as estimativas dos coeficientes, das predições e as covariâncias das próprias estimativas. Além disso, uma discussão mais aprofundada das implicações econômico-sociais que os resultados desse estudo podem fornecer.

Já havia sinalizado na seção dos resultados sobre a ausência do último parágrafo da Seção 4. Ficarei no aguardo.

A Oceania foi excluída de parte das análises, mas essa informação não aparece no texto.

Uma última sugestão: em modelagem é usual, após a escolha do modelo, traçarmos um ou dois perfis com base no ajuste. A ideia é informar os valores/categorias de um país (no caso de vocês) e informar qual seria o y estimado.

Parabéns pelo texto. Os relatórios de vocês são sempre muito bem apresentados e são os únicos que incluíram as referências.

# REFERÊNCIAS


Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, **160**(901), 268–282.

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.

Chancel, L., Piketty, T., Saez, E. e Zucman, G. (2022). *World inequality report 2022*. Harvard University Press.

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.

Hoffmann, R. (2016). *Análise de regressão: uma introdução à econometria*. Portal de Livros Abertos da USP.

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.