Trabalho 2

SME0620 – Estatística

Prof. Dr. Vicente Garibay Cancho

Integrantes:

Daniel Umeda Kuhn – 13676541

Manoel Thomaz Gama da Silva Neto - 13676392

São Carlos – SP

2024

Sumário

1. Introdução

Neste relatório iremos realizar uma análise de indicadores socioeconômicos dos países do mundo, utilizando análise de regressão, análise de ANOVA e teste qui-quadrado, de modo a observar como os dados se comportam quando submetidos a diferentes testes, e averiguar possíveis correlações entre eles.

2. Metodologia

Nesta seção, detalharemos os métodos estatísticos aplicados na análise dos dados do estudo, focando em três técnicas principais: análise de regressão, análise de variância (ANOVA) e teste qui-quadrado. Para estudar as relações entre os dados, aplicaremos essas técnicas de análise por meio da linguagem de programação R, relacionando variáveis tanto quantitativas quanto qualitativas.

2.1. Análise de Regressão

A análise de regressão foi utilizada para explorar a relação entre uma variável dependente e uma ou mais variáveis independentes. Esta técnica é crucial para previsões e para compreender como as variáveis independentes afetam a variável dependente. Foi aplicado 1 tipo de regressão: regressão linear simples, mas a título de curiosidade também detalhamos o funcionamento de outras outras 3 análises: a regressão linear múltipla, a regressão polinomial, e a regressão logística.

2.1.1. Regressão Linear Simples

Utilizada quando há apenas uma variável independente, o modelo de regressão linear simples é representado pela equação:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Onde Y é a variável dependente, X é a variável independente, β_0 é o intercepto, β_1 é o coeficiente de inclinação, e ε é o termo de erro. Os parâmetros β_0 e β_1 são estimados pelo método dos mínimos quadrados, que minimiza a soma dos quadrados das diferenças entre os valores observados

e os previstos de *Y*. A adequação do modelo foi verificada pelo coeficiente de determinação (R^2).

2.1.2. Regressão Linear Múltipla

Quando várias variáveis independentes estavam presentes, a regressão linear múltipla foi utilizada, descrita pela equação:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Os parâmetros foram estimados pelo método dos mínimos quadrados generalizados. A significância dos coeficientes individuais foi avaliada com testes t, e a significância global do modelo foi verificada com o teste F. O coeficiente de determinação ajustado (R^2 ajustado) foi utilizado para considerar o número de variáveis independentes no modelo.

2.1.3. Regressão Polinomial

Quando a relação entre a variável dependente e a variável independente não é linear, a regressão polinomial é aplicada. O modelo é uma extensão da regressão linear e é descrito pela equação:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n + \varepsilon$$

Os coeficientes são estimados de maneira semelhante à regressão linear, mas considerando os termos polinomiais.

2.1.4. Regressão Logística

A regressão logística é utilizada quando a variável dependente é categórica, geralmente binária (0 ou 1). O modelo de regressão logística é dado por:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

onde p é a probabilidade de um evento ocorrer. A função logística transforma a relação linear entre as variáveis independentes e a variável dependente em uma relação probabilística.

2.2. Análise de Variância (ANOVA)

A ANOVA foi empregada para comparar as médias de três ou mais grupos e identificar diferenças estatisticamente significativas entre eles. Utilizamos a ANOVA de um fator, que avalia um único fator ou variável categórica.

Procedimentos da ANOVA:

Hipóteses:

H_0: As médias dos grupos são iguais.

H_1: Pelo menos uma média de grupo é diferente.

Cálculo das Somatórias:

Soma dos Quadrados entre os Grupos (*SSB*): Variabilidade entre as médias dos grupos.

Soma dos Quadrados dentro dos Grupos (*SSW*): Variabilidade dentro de cada grupo.

Soma Total dos Quadrados (*SST*): Soma de *SSB* e *SSW*.

Graus de Liberdade:

Entre os grupos: $df_B = k - 1$, onde k é o número de grupos.

Dentro dos grupos: $df_W = N - k$, onde N é o total de observações.

Total: $df_T = N - 1$

Cálculo das Médias dos Quadrados:

MSB = SSB / df_B

MSW = SSW / df_W

Estatística F:

F = MSB / MSW

Decisão: Comparação do valor de *F* com o valor crítico da distribuição *F*. Se *F* calculado for maior que *F* crítico, rejeita-se H_0.

2.3. Teste Qui-Quadrado

Os testes qui-quadrado são ferramentas estatísticas amplamente utilizadas para analisar dados categóricos. Eles ajudam a determinar se há associações entre variáveis categóricas ou se uma distribuição observada segue uma distribuição teórica esperada. Existem três tipos principais de testes qui-quadrado: teste de aderência, teste de independência e teste de homogeneidade, que serão explicados a seguir.

2.3.1. Teste de Aderência

O teste de aderência verifica se uma distribuição observada segue uma distribuição teórica específica. É utilizado para avaliar se os dados observados em diferentes categorias se ajustam a uma distribuição esperada.

Passos para Realizar o Teste de Aderência:

Formular as Hipóteses:

Hipótese nula (H_0): Os dados seguem a distribuição teórica esperada.

Hipótese alternativa (H_1): Os dados não seguem a distribuição teórica esperada.

Determinar as Frequências Esperadas: As frequências esperadas (E_i) são calculadas com base na distribuição teórica. Por exemplo, se a

distribuição teórica for uniforme, as frequências esperadas serão iguais para todas as categorias.

Calcular a Estatística Qui-Quadrado: A estatística qui-quadrado (χ^2) é calculada pela fórmula:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

onde O_i são as frequências observadas e E_i são as frequências esperadas.

Determinar os Graus de Liberdade: Os graus de liberdade (df) para o teste de aderência são:

$$df = k - 1 - m$$

Onde k é o número de categorias e m é o número de parâmetros estimados da distribuição teórica.

Comparar com o Valor Crítico: Compare a estatística χ^2 calculada com o valor crítico da distribuição qui-quadrado para os graus de liberdade e nível de significância (α\alphaα) escolhidos. Se χ^2 calculado > χ^2 crítico, rejeitamos H_0.

2.3.2. Teste qui-quadrado de independência

O teste de independência é utilizado para determinar se há uma associação significativa entre duas variáveis categóricas. É usado principalmente em tabelas de contingência.

Passos para Realizar o Teste de Independência:

Formular as Hipóteses:

Hipótese nula (H_0): As variáveis são independentes.

Hipótese alternativa (H_1): As variáveis não são independentes.

Construir a Tabela de Contingência: A tabela de contingência apresenta as frequências observadas para cada combinação de categorias das duas variáveis.

Calcular as Frequências Esperadas: As frequências esperadas (E_{ij}) são calculadas sob a hipótese de independência:

$$E_{ij} = \frac{R_i \times C_j}{N}$$

onde R_i é o total da linha i, C_j é o total da coluna jjj e NNN é o total geral de observações.

Calcular a Estatística Qui-Quadrado: A estatística qui-quadrado (χ^2) é calculada pela fórmula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

onde O_{ij} são as frequências observadas e E_{ij} são as frequências esperadas.

Determinar os Graus de Liberdade: Os graus de liberdade (*df*) para o teste de independência são:

$$df = (r - 1)(c - 1)$$

onde r é o número de linhas e c é o número de colunas na tabela de contingência.

Comparar com o Valor Crítico: Compare a estatística χ^2 calculada com o valor crítico da distribuição qui-quadrado para os graus de liberdade e nível de significância (α) escolhidos. Se $\chi^2_{\text{calculado}} > \chi^2_{\text{crítico}}$, rejeitamos H_0.

2.3.3. Teste qui-quadrado de homogeneidade

O teste de homogeneidade é usado para verificar se duas ou mais populações têm a mesma distribuição de uma variável categórica. Ele é semelhante ao teste de independência, mas é aplicado em diferentes populações.

Passos para Realizar o Teste de Homogeneidade:

Formular as Hipóteses:

Hipótese nula (H_0): As populações têm a mesma distribuição da variável categórica.

Hipótese alternativa (H_1): As populações têm distribuições diferentes da variável categórica.

Construir a Tabela de Contingência: A tabela de contingência apresenta as frequências observadas para cada combinação de categorias e populações.

Calcular as Frequências Esperadas: As frequências esperadas (E_{ij}) são calculadas da mesma forma que no teste de independência:

$$E_{ij} = \frac{R_i \times C_j}{N}$$

Calcular a Estatística Qui-Quadrado: A estatística qui-quadrado (χ^2) é calculada pela fórmula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Determinar os Graus de Liberdade: Os graus de liberdade (*df*) são:

$$df = (r - 1)(c - 1)$$

Comparar com o Valor Crítico: Compare a estatística χ2\chi^2χ2 calculada com o valor crítico da distribuição qui-quadrado para os graus de liberdade e nível de significância (α\alphaα) escolhidos. Se $\chi^2_{\text{calculado}} > \chi^2_{\text{crítico}}$, rejeitamos H_0.

3. Resultados

3.1 Teste da regressão linear

Nesta seção apresentamos os gráficos relativos aos testes efetuados descritos na seção 2.

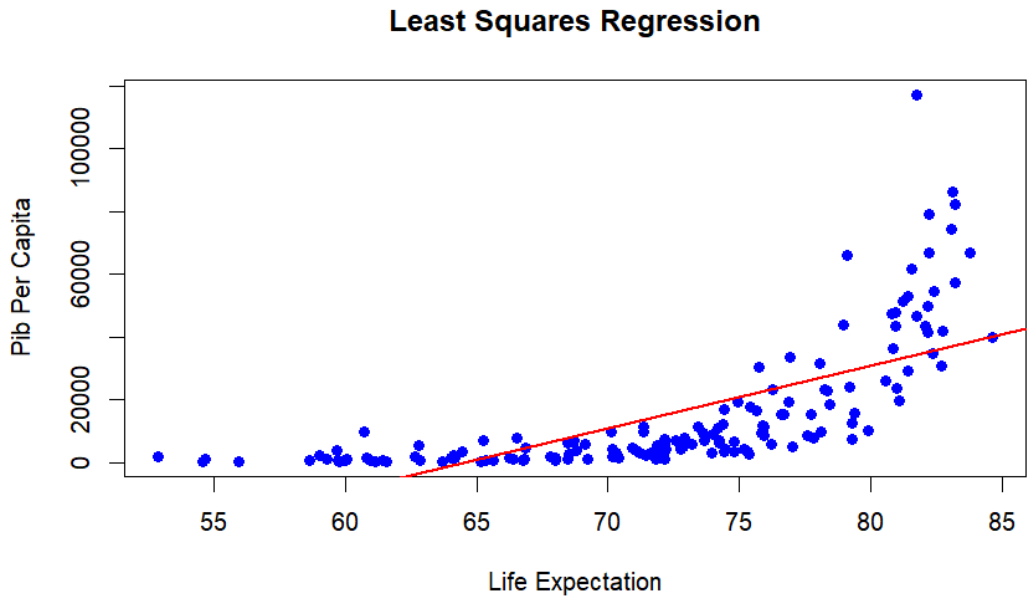

Gráfico 1: Análise de Regressão Linear Simples(mínimos quadrados)

O eixo x representa a expectativa de vida e o eixo y representa o pib per capita de um dado país. O gráfico foi obtido a partir da regressão linear simples a partir do método dos mínimos quadrados. Observa-se razoável correlação linear, visto que a maioria dos pontos encontra-se próximo à linha de regressão.

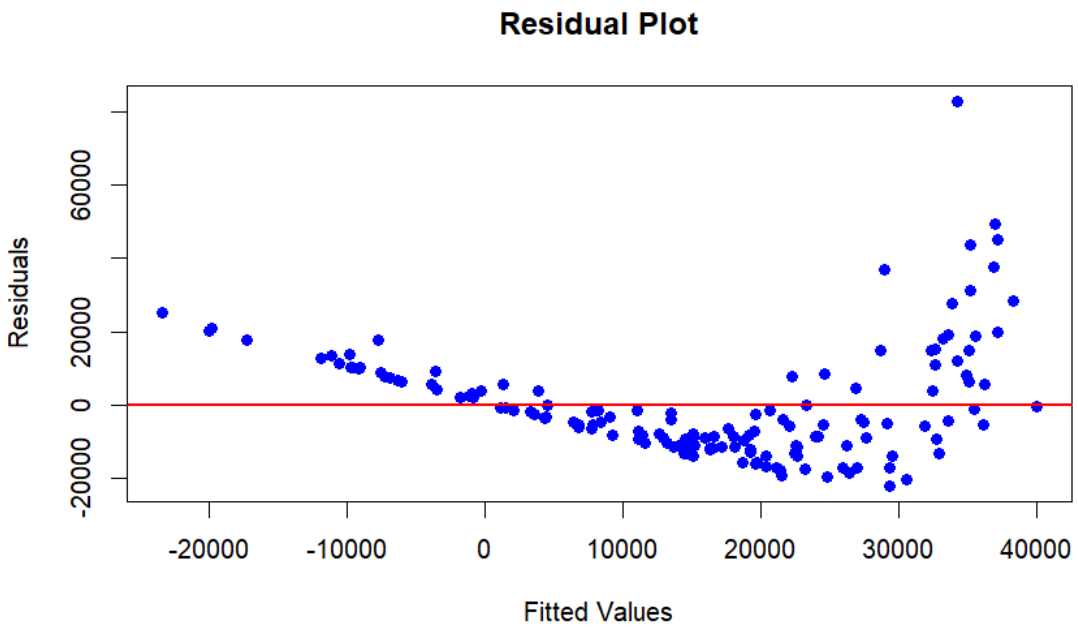

Gráfico 2: Resíduos da Regressão Linear Simples

O gráfico residual da regressão linear simples demonstra que a maior parte dos pontos encontram-se distribuídos em torno da reta $y = 0$. Isso é sinal de uma boa amostra para esse tipo de análise.

```
[1] "Structure of the data:"
'data.frame':   150 obs. of  5 variables:
 $ Code                     : chr  "ABW" "ALB" "ARE" "ARG" ...
 $ LifeExpec                : num  75.7 77 78.9 75.9 72.2 ...
 $ FormalEducatedPopulation: num  94 98 93 98 99 99 100 99 63 97 ...
 $ IDHRank                  : num  26 4 15 38 8 8 34 178 22 179 ...
 $ PibPerCapita             : num  30253 5288 43839 11795 4221 ...
[1] "Summary of the least squares model:"

Call:
lm(formula = PibPerCapita ~ LifeExpec, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-22078 -10812  -4067   7818  82888

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -129165.4    12615.6  -10.24   <2e-16 ***
LifeExpec      1999.9      173.1   11.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15560 on 148 degrees of freedom
Multiple R-squared:  0.4743,    Adjusted R-squared:  0.4707
F-statistic: 133.5 on 1 and 148 DF,  p-value: < 2.2e-16
```

Gráfico 3: Estrutura da Regressão Linear Simples pelos comandos “str” e “summary”

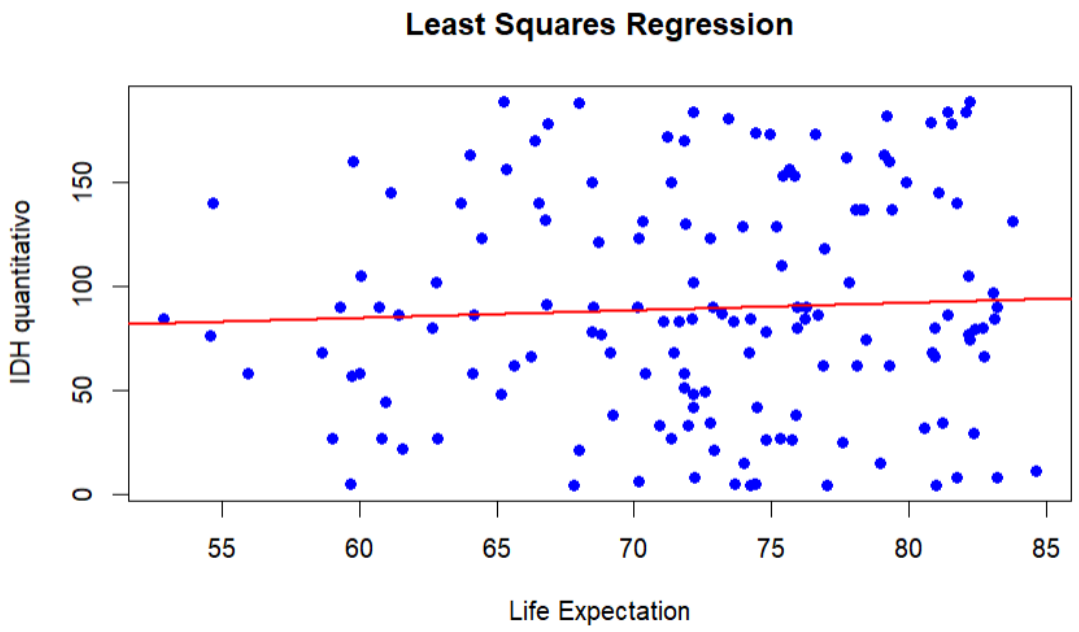

Gráfico 4: Análise de Regressão Linear Simples(mínimos quadrados)

O eixo x representa a expectativa de vida e o eixo y representa o IDH quantitativo de um dado país. O gráfico foi obtido a partir da regressão linear simples a partir do método dos mínimos quadrados. Observa-se uma péssima correlação linear, visto que a maioria dos pontos encontram-se dispersos em relação à linha de regressão.

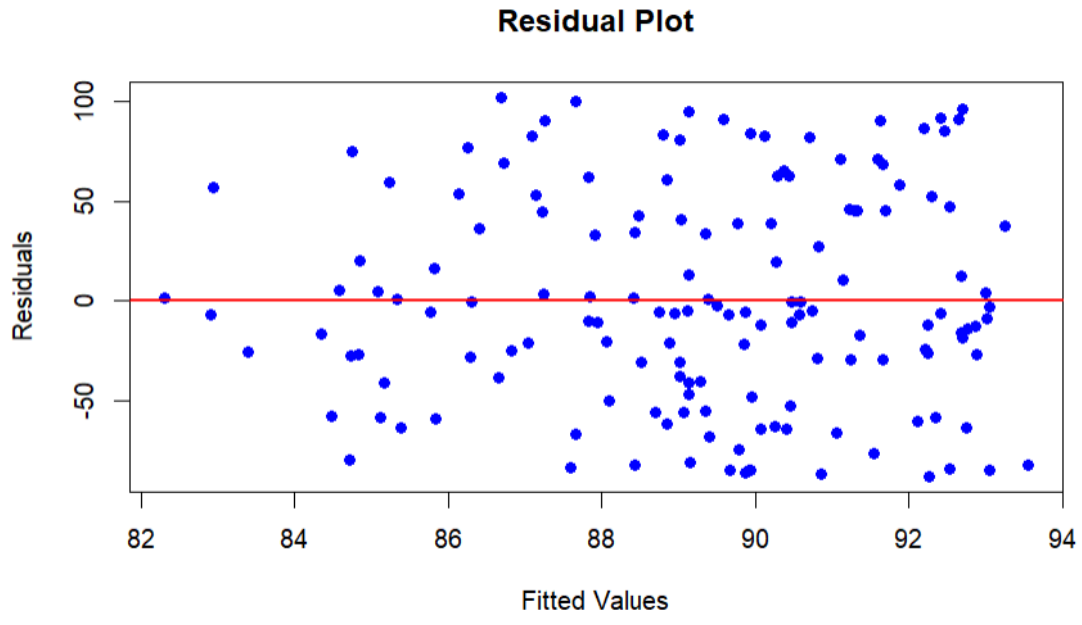

Gráfico 5: Resíduos da Regressão Linear Simples

O gráfico residual da regressão linear simples demonstra que a maior parte dos pontos encontram-se dispersos com relação à reta $y = 0$. Isso é sinal de uma amostra não adequada para esse tipo de análise.

```
[1] "Summary of the least squares model:"

Call:
lm(formula = IDHRank ~ LifeExpec, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-88.268 -40.916  -5.909  45.574 102.311

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.5380    43.8808   1.448    0.150
LifeExpec     0.3548     0.6020   0.589    0.557

Residual standard error: 54.11 on 148 degrees of freedom
Multiple R-squared:  0.002341,  Adjusted R-squared:  -0.0044
F-statistic: 0.3473 on 1 and 148 DF,  p-value: 0.5565
```

Gráfico 6: Estrutura da Regressão Linear Simples pelos comandos "str" e "summary"

3.2 Teste ANOVA

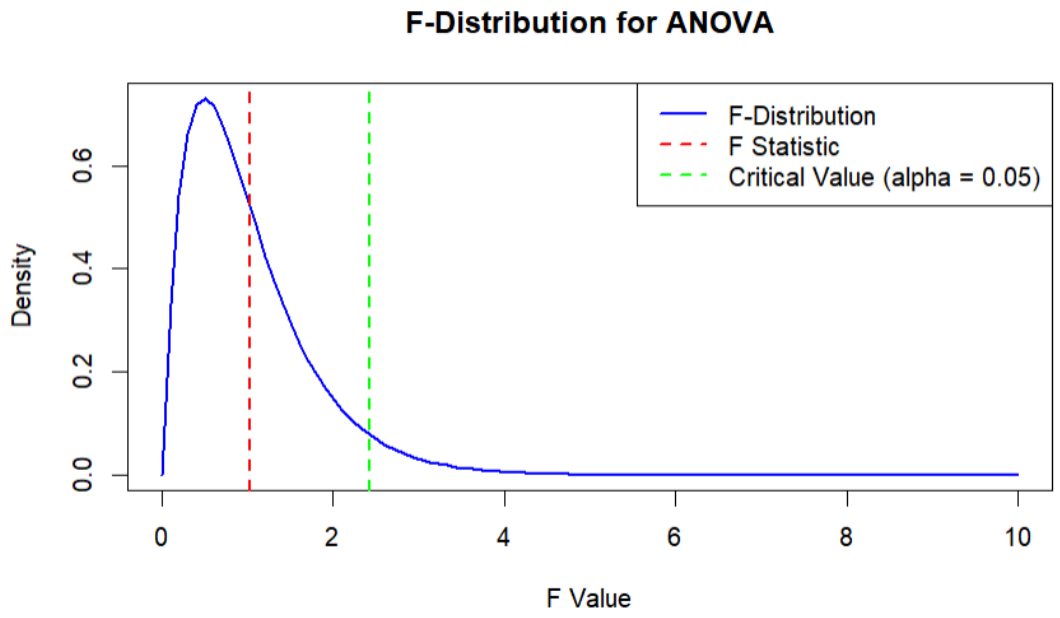

Gráfico 7: Gráfico da distribuição F para o teste ANOVA

A partir do gráfico da distribuição F podemos perceber que há uma semelhança significativa na média dos dados pela posição relativa da linha crítica à direita da estatística F observada, indicativo de uma aceitação da hipótese nula

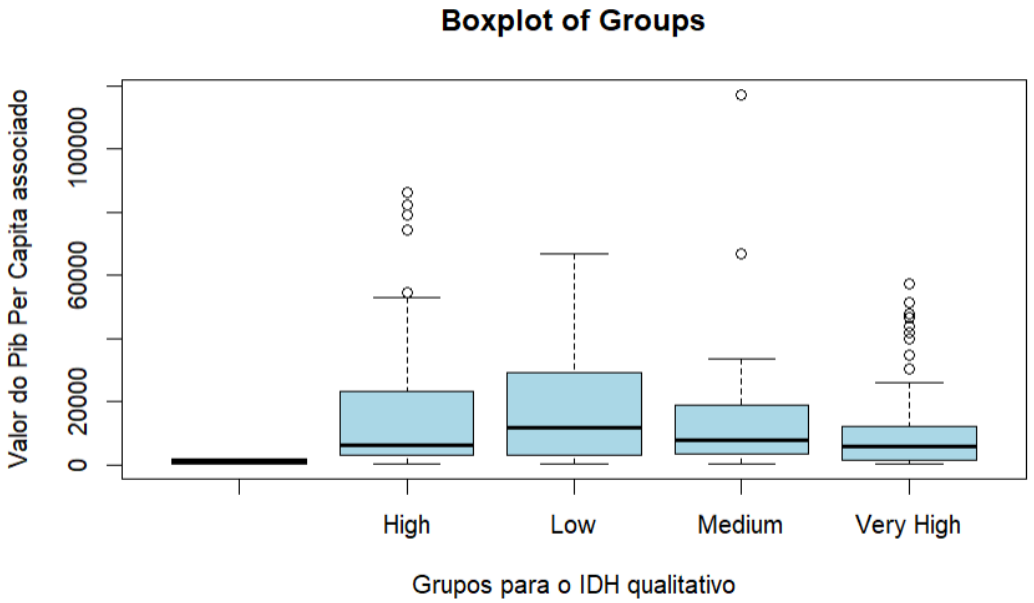

Gráfico 8: Gráfico de BoxPlot dos grupos utilizados versus Pib Per Capita

Podemos perceber pela posição das caixas que os grupos de IDH não estão tão bem associados com o Pib Per Capita no sentido de que os grupos com melhores indicadores sociais de IDH não necessariamente apresentam as faixas de maior Pib Per Capita .

Obs.: o grupo da esquerda representa os outliers que são países que devido ao modelo político ou à condições do contexto social não foram listados nesses dados.

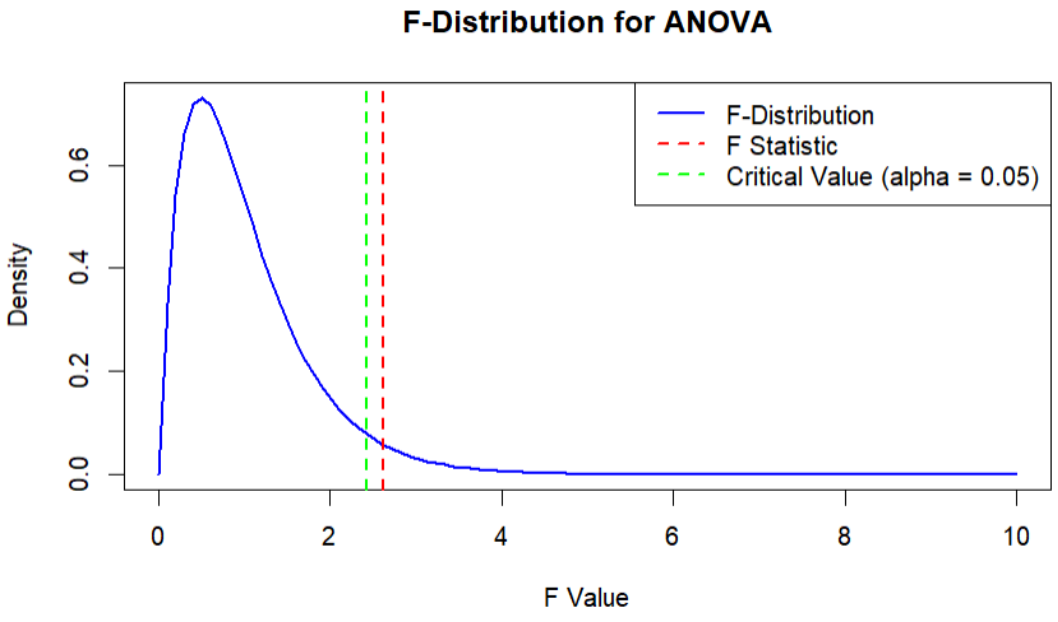

Gráfico 9: Gráfico da distribuição F para o teste ANOVA

A partir do gráfico da distribuição F podemos perceber que há uma diferença significativa na média dos dados pela posição relativa da linha crítica à esquerda da estatística F observada, indicativo de uma rejeição da hipótese nula.

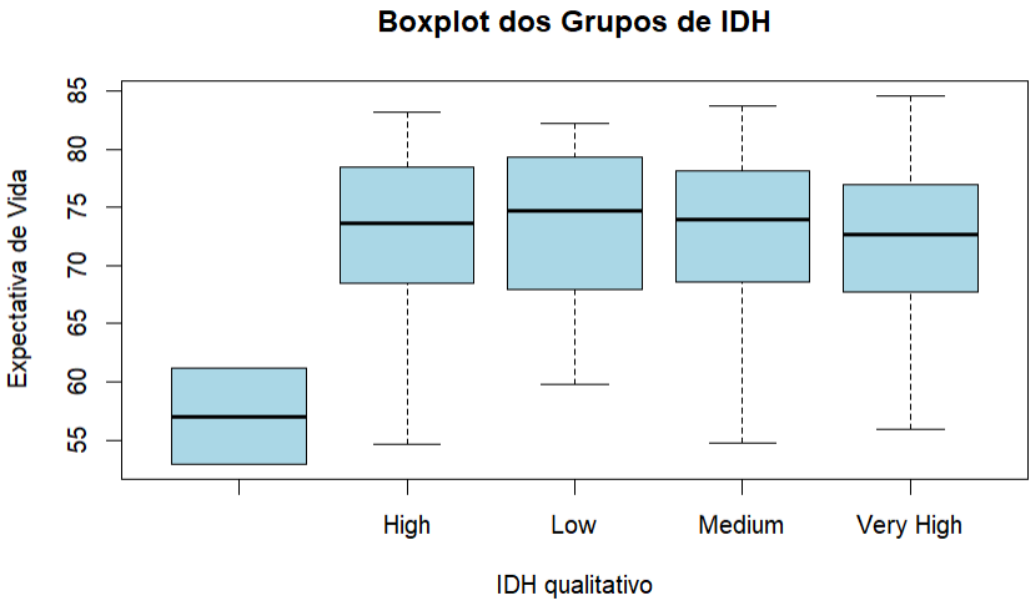

Gráfico 10: Gráfico de BoxPlot dos grupos utilizados versus Pib Per Capita

Podemos perceber pela posição das caixas que os grupos de IDH apresentam uma distribuição média de expectativa de vida próxima independentemente dos grupos em que se encontram, esse fato pode ser explicado com o fato de que a expectativa tem diversos fatores genéticos que podem influir variando de maneira muito aleatória.

Obs.: o grupo da esquerda representa os outliers que são países que devido ao modelo político ou à condições do contexto social não foram listados nesses dados

3.3 Teste do Qui-Quadrado

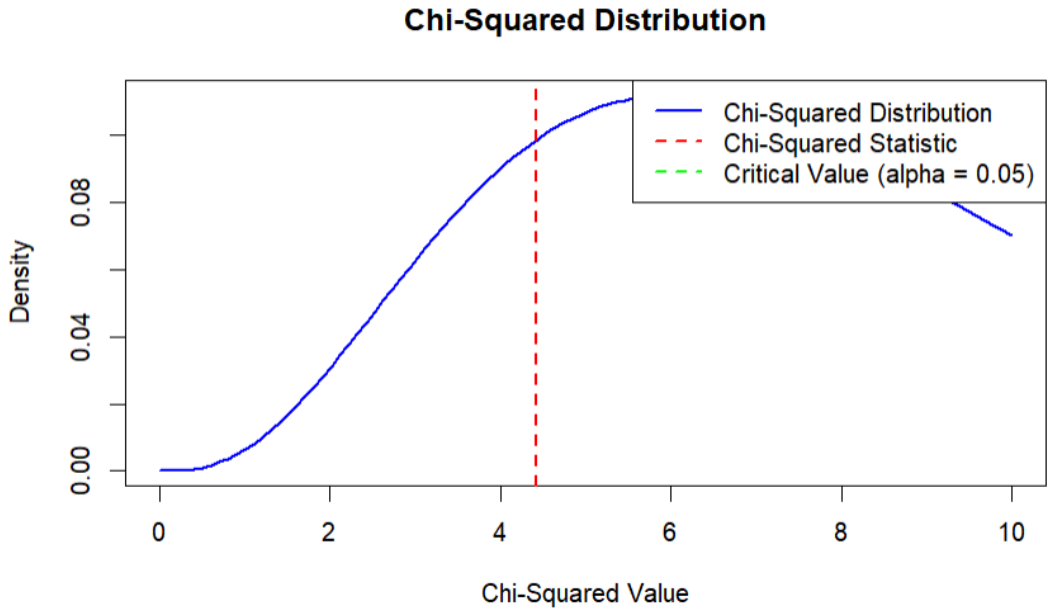

Gráfico 10: Gráfico da distribuição Chi-Quadrado

No Gráfico acima a linha do valor crítico está localizada à direita do gráfico estando assim a direita do valor de Chi medido, sendo assim dando forte evidência de que a hipótese nula não pode ser negada, ou seja, os dados seriam independentes.

Obs.: a distância entre a linha de significância e o valor medido sugere que a base de dados não é adequada para o uso do método acima sendo necessário o uso de um método como a análise exata de Fischer.

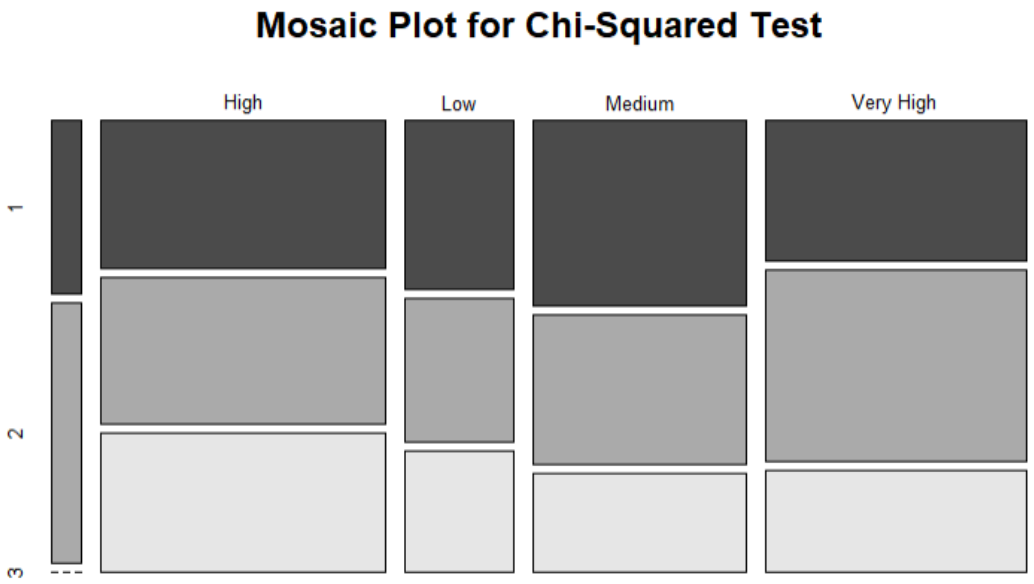

Gráfico 11: Gráfico da distribuição em mosaico dos dados.

No Gráfico acima podemos ver como os dados de primeiro,segundo e terceiro mundo estão distribuídos versus as classificações em indicadores sociais.

```
[1] "Contingency table:"

             1  2  3
             2  3  0
High        16 16 15
Low          7  6  5
Medium      15 12  8
Very High   14 19 10
```

Gráfico 12: Tabela de contingência .

```
[1] "Chi-squared test result:"

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 4.4264, df = 8, p-value = 0.8168
```

Gráfico 13: Resultados do teste chi-quadrado.

Assim, é possível inferir dos resultados obtidos que, ao analisar o arranjo do gráfico, nossa hipótese estava correta, e os dados coletados não são adequados para a aplicação do método qui-quadrado, o que pode gerar um erro do tipo II.

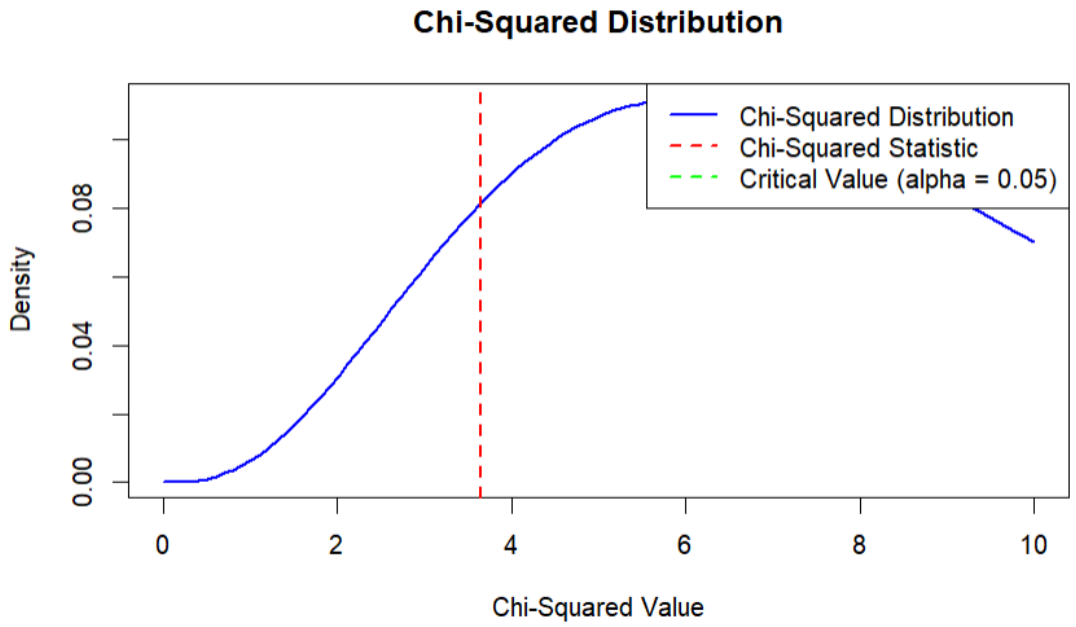

Gráfico 14:Gráfico da distribuição Chi-Quadrado

No Gráfico acima a linha do valor crítico está localizada à direita do gráfico estando assim a direita do valor de Chi medido, sendo assim dando forte evidência de que a hipótese nula não pode ser negada, ou seja, os dados seriam independentes.

Obs.: a distância entre a linha de significância e o valor medido sugere que a base de dados não é adequada para o uso do método acima sendo necessário o uso de um método como a análise exata de Fischer.

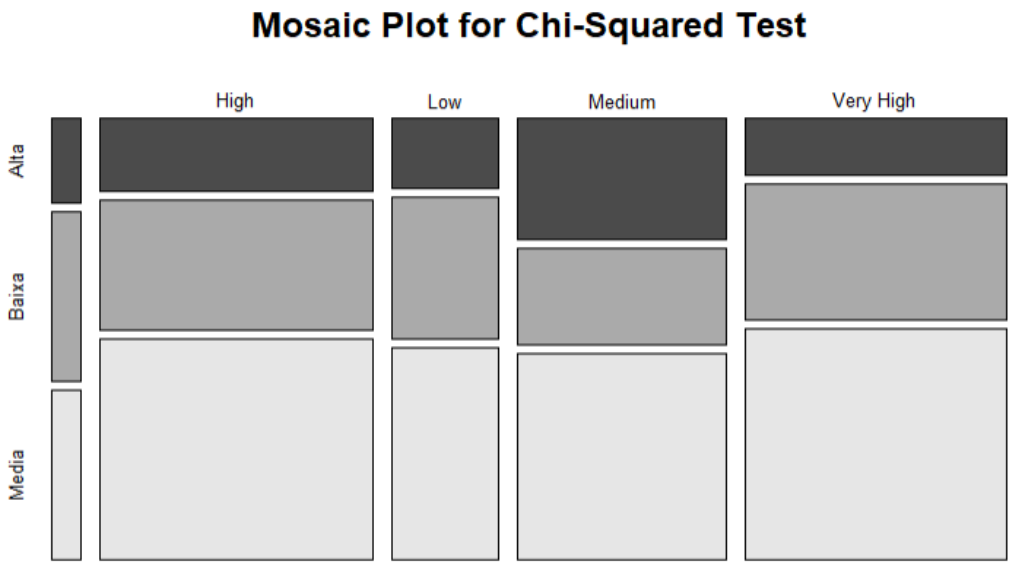

Gráfico 15: Gráfico da distribuição em mosaico dos dados.

```
[1] "Contingency table:"

            Alta Baixa Media
               1     2     2
  High         8    14    24
  Low          3     6     9
  Medium      10     8    17
  Very High    6    14    24
```

Gráfico 16: Tabela de contingência .

```
[1] "Chi-squared test result:"

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 4.4264, df = 8, p-value = 0.8168
```

Gráfico 17: Resultados do teste chi-quadrado.

Por fim, podemos perceber pelos resultados obtidos que a nossa hipótese ao analisar a disposição do gráfico estava correta e os dados não são apropriados para uso do método qui-quadrado podendo ocasionar um erro do tipo II.

4. Conclusão

Após examinar os resultados, podemos perceber que os ambos os dados de Pib Per Capita e IDH e indicadores sociais apresentam alguma correlação entre si. Ao final do primeiro teste da regressão linear podemos perceber que a base de dados do Pib Per Capita se adequa melhor a esse teste demonstrando uma correlação com a expectativa de vida, por outro lado o IDH quantitativo apresenta dados esparsos, fator que contribui com a ineficiência dessa análise para a base de dados escolhida na pesquisa. Já no teste ANOVA podemos perceber pela proximidade do valor crítico e o valor observado que os dados apresentam uma certa correlação e apontam uma correlação levemente maior para os fatores e indicativos sociais em detrimento do Pib Per capita. Já na última análise, utilizando o método do chi-quadrado podemos perceber que não é adequada para nenhuma das duas bases de dados visto que para tal o mesmo deve apresentar um valor de p menor ou próximo de 0,05 para ser confiável.

Desse modo, as análises em geral contribuem para a hipótese formulada ao longo da pesquisa, de que fatores sociais são afetados mais por outros fatores sociais do que por fatores econômicos. Assim, recomenda-se por meio da examinação dos dados, que haja maior atenção e investimentos por parte dos governos dos países preponderantemente nos quesitos sociais, como educação e saúde pública, pois estes refletem melhor na qualidade de vida e bem-estar social das populações.