ANÁLISE MULTIVARIADA DOS FATORES SOCIOECONÔMICOS E INDICADORES DE SAÚDE PÚBLICA: UM ESTUDO COM O CONJUNTO DE DADOS DE EXPECTATIVA DE VIDA DA OMS

ME731 - MÉTODOS EM ANÁLISE MULTIVARIADA

Bianca Aires de Sousa - RA 204223

Novembro de 2024

1 Introdução

O presente trabalho busca analisar um conjunto de dados com o propósito de identificar o modelo estatístico multivariado mais adequado, que capture as relações entre as variáveis de interesse. O banco de dados selecionado fornece informações relevantes sobre indicadores de saúde, econômicos e sociais, permitindo uma análise aprofundada de como essas variáveis se relacionam entre si. A partir disso, busca-se entender quais fatores possuem maior impacto sobre variáveis de desfecho, como a expectativa de vida e a mortalidade, e assim fornecer uma interpretação significativa dos dados.

2 Metodologia

2.1 Análise Descritiva

Com o objetivo de estudar o comportamento das variáveis individualmente, foi realizada uma análise descritiva dos dados, a qual utilizou o software estatístico *RStudio* para a elaboração de gráficos, tabelas e cálculo de estatísticas descritivas, como média, mediana, desvio padrão e frequência, que ajudam a descrever o comportamento dos dados e as relações entre as variáveis.

2.2 Regressão Múltipla Multivariada

A Regressão Múltipla Multivariada é uma abordagem estatística que permite analisar simultaneamente diversas variáveis dependentes, utilizando um conjunto de variáveis independentes como base.

O modelo estatístico para uma análise multivariada com m variáveis dependentes e n observações é expresso como:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

onde:

- **Y** é a matriz n x m das variáveis dependentes;
- **X** é a matriz n x p, onde (p-1) corresponde ao número de variáveis independentes;
- **B** é a matriz p x m dos coeficientes a serem estimados;
- **E** é a matriz n x m dos erros.

A validação do modelo depende das seguintes suposições: normalidade multivariada dos resíduos (com média zero e variância constante para todas as variáveis independentes), e independência entre os resíduos.

Após a definição do modelo de regressão, realizam-se testes de hipóteses para os coeficientes do modelo, a fim de verificar a significância das variáveis preditoras em relação a cada resposta.

2.3 Análise de Variância Multivariada (MANOVA)

A aplicação da Análise de Variância Multivariada (MANOVA), após a análise de regressão, é útil para avaliar se as variáveis preditoras têm um efeito estatisticamente significativo sobre o conjunto das variáveis dependentes.

3 Resultados

3.1 Análise Descritiva e Exploratória

O banco de dados escolhido para as análises a seguir chama-se *Life Expectancy (WHO)*, disponível na plataforma Kaggle. O banco reúne informações de saúde e socioeconômicas de vários países no período de 2000 a 2015, que foram disponibilizados pela Organização Mundial de Saúde (OMS), e tem ao todo 1649 observações, após a retirada de dados faltantes.

Dentre as 22 variáveis totais, as seguintes variáveis foram selecionadas para serem analisadas por questão contextual:

- *Life Expectancy*: Expectativa de vida — reflete o resultado geral de saúde da população.
- *Adult Mortality*: Mortalidade adulta — indica o índice de mortalidade na população adulta, ligado à saúde e condições de vida.

- *Alcohol*: Consumo de álcool - representa o consumo per capita de álcool em litros.
- *BMI*: Índice de Massa Corporal — indica o nível médio de saúde nutricional da população.
- *GDP*: PIB per capita — fator econômico que pode influenciar amplamente as condições de saúde.
- *Schooling*: Escolaridade — representa o nível de educação, que está ligado a melhores práticas de saúde e acesso a serviços.

Para entender melhor o comportamento dos baixos, foram calculadas medidas descritivas das variáveis, que podem ser encontradas na tabela abaixo:

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Expectativa de Vida	44.0	64.4	71.7	69.3	75.0	89.0
Mortalidade Adulta	1.0	77.0	148.0	168.2	227.0	723.0
Álcool	0.010	0.810	3.790	4.533	7.340	17.870
IMC	2.0	19.5	43.7	38.13	55.8	77.1
Escolaridade	4.20	10.30	12.30	12.12	14.00	20.70
PIB	1.68	462.15	1592.57	5566.03	4718.51	119172.74
Log(PIB)	0.52	6.136	7.373	7.316	8.46	11.688

Tabela 1: Medidas Descritivas do Conjunto de Dados

Além disso, as distribuições das variáveis também podem ser úteis para analisar seu comportamento. Assim, a seguir estão os histogramas de cada variável:

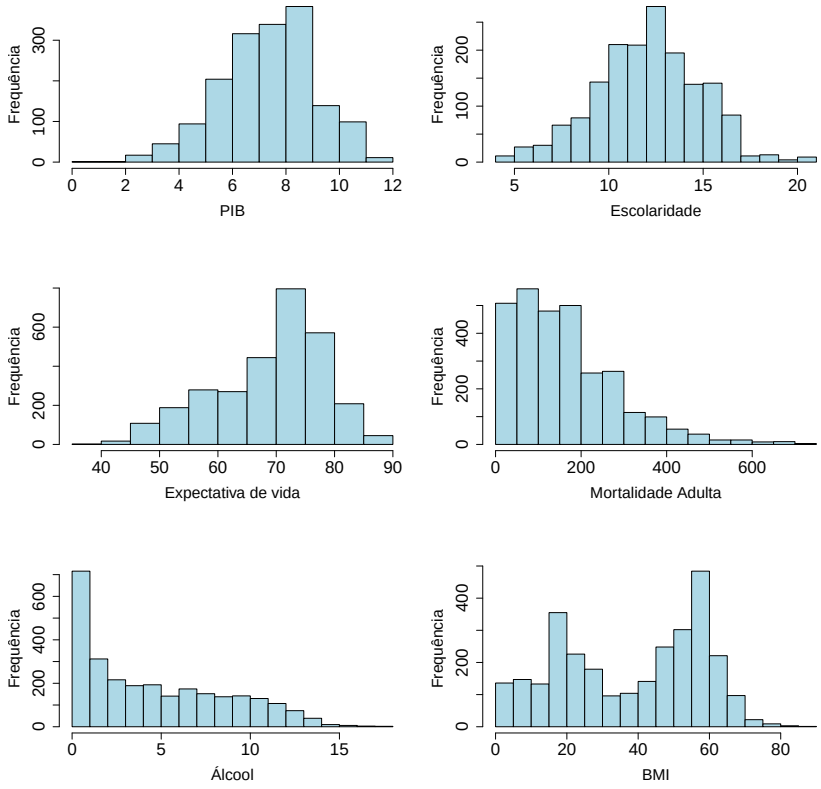

Figura 1: Distribuição das variáveis

A expectativa de vida varia de 44 a 89 anos, com uma média de 69,3 anos. Isso indica uma ampla variação entre os países, refletindo diferentes condições de saúde e qualidade de vida. A mediana de 71,7 anos e o terceiro quartil de 75,0 anos mostram que metade dos países têm expectativa de vida entre esses valores, sugerindo que a maior parte dos países tem uma expectativa de vida em torno de 70 anos. Isso pode ser observado no histograma, que mostra uma distribuição aproximadamente simétrica e levemente concentrada em torno dos valores centrais

(próximo a 70 anos).

A mortalidade adulta varia bastante, de 1 até 723 por 1000 adultos, com uma média de 168,2. Isso reflete disparidades significativas nas condições de saúde e fatores de risco entre os países. A mediana de 148,0 e o terceiro quartil de 227 indicam que a maioria dos países tem níveis de mortalidade adulta mais baixos, mas alguns países têm valores muito altos. No histograma, é possível observar essa assimetria à direita, com a maior parte dos dados concentrada em valores mais baixos e uma cauda longa em direção aos valores mais altos.

O consumo de álcool per capita varia de 0,010 a 17,87 litros, com uma média de 4,533. A mediana de 3,79 litros indica que metade dos países consome até esse valor, sugerindo que o consumo é moderado na maioria dos casos, com alguns países apresentando consumo bem mais elevado. Essa assimetria à direita também pode ser observada no histograma.

O IMC (BMI) médio é 38,3, com uma variação de 2 a 77,1. A mediana de 43,7 e o terceiro quartil de 55,8 mostram que o IMC é relativamente alto, indicando possíveis problemas de obesidade em muitos países. No histograma, observa-se dois "picos"na distribuição, próximo a 20 e 60.

A escolaridade varia entre 4,2 e 20,7 anos, com uma média de 12,12 anos. Isso sugere que há países com baixa escolaridade média e outros com níveis mais altos. A mediana de 12 anos indica que metade dos países tem uma média de escolaridade até esse valor. O histograma mostra uma distribuição simétrica em torno dos valores centrais entre 10 e 15 anos.

O PIB per capita varia significativamente, de 1,68 até 119.172,74, com uma média de 5566,03. Essa grande variação reflete as diferenças econômicas entre os países. Para ajustar essa alta variação e facilitar a análise multivariada, tomou-se o logaritmo do PIB para as análises a seguir, a fim de transformar esses dados em uma escala mais comparável e linearizar a relação entre o PIB e as variáveis de saúde. Isso pode tornar a interpretação dos resultados mais intuitiva.

Para estudar as relações entre as variáveis, a matriz abaixo mostra os gráficos de dispersão entre as variáveis, bem como a correlação entre elas:

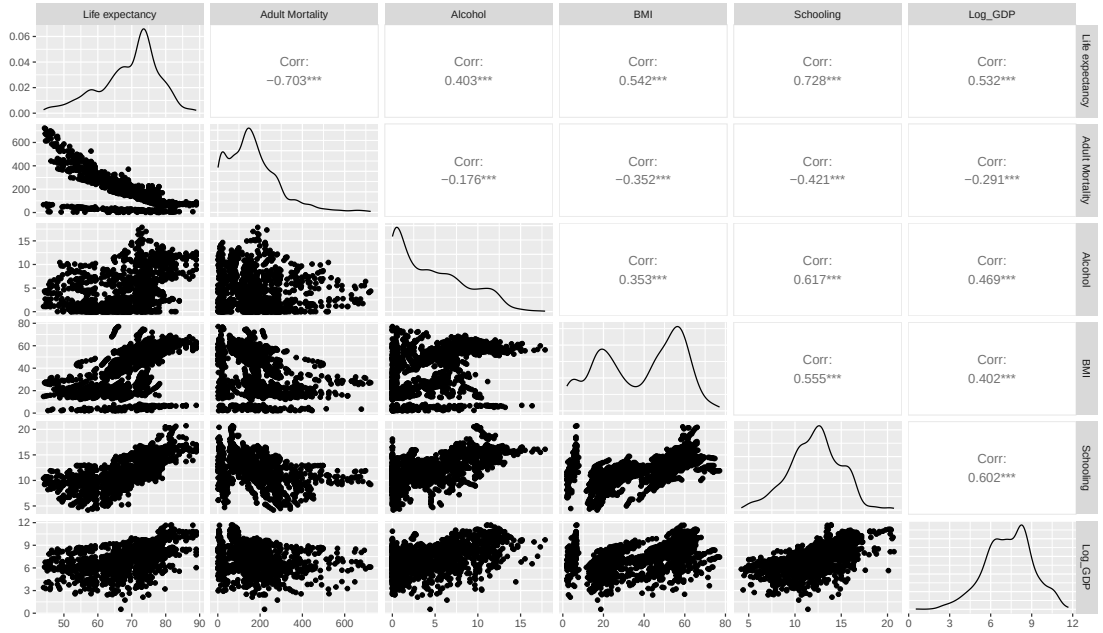

Figura 2: Matriz de gráficos de dispersão

Pode-se observar que expectativa de vida apresenta correlação positiva forte com escolaridade e moderada com o logaritmo do PIB, indicando que, em países com maior nível de educação e riqueza, a população tende a viver mais. A relação entre expectativa de vida e essas variáveis socioeconômicas parece ser linear, o que sugere que um modelo de regressão multivariada pode capturar bem essas associações. A mortalidade adulta, por outro lado, apresenta uma correlação negativa com escolaridade e expectativa de vida, coerente com a ideia de que educação e saúde pública ajudam a reduzir a mortalidade.

Outras variáveis, como consumo de álcool e IMC, também estão relacionadas de forma moderada e positiva, indicando uma possível associação entre estilo de vida e hábitos de saúde nos países analisados. A escolaridade e

o PIB mostram uma correlação positiva, reforçando a relação esperada entre riqueza e acesso à educação.

Dada a estrutura correlacionada dos dados e as tendências lineares observadas, a regressão multivariada é o método ideal para explorar simultaneamente o impacto das variáveis socioeconômicas, como escolaridade e PIB, sobre múltiplos indicadores de saúde e bem-estar, como expectativa de vida, mortalidade adulta, consumo de álcool e IMC. Esse modelo permite não só avaliar o efeito individual de cada preditora sobre as variáveis-resposta, mas também entender o impacto conjunto dessas variáveis socioeconômicas. Em resumo, a regressão multivariada oferece uma abordagem robusta para investigar como fatores como educação e riqueza afetam aspectos inter-relacionados da saúde pública e qualidade de vida.

3.2 Modelo de Regressão Multivariada

Para essa análise, as variáveis-resposta escolhidas foram Expectativa de Vida, Mortalidade Adulta, Consumo de Álcool e IMC, pois representam diferentes aspectos de saúde e bem-estar. As variáveis-preditoras selecionadas foram Escolaridade e PIB (na forma logarítmica), como principais indicadores socioeconômicos que possivelmente influenciam essas condições de saúde.

Para a realização dos modelos de regressão múltipla multivariada, foi utilizada a função *lm* disponível no R. Os resultados obtidos podem ser visualizados a seguir:

	Estimativa	Erro Padrão	Valor t	Valor p
(Intercepto)	39.51003	0.71243	55.458	< 2e-16 ***
Log_PIB	0.74361	0.10474	7.099	1.86e-12 ***
Escolaridade	2.00923	0.06564	30.610	< 2e-16 ***

Tabela 2: Modelo para Expectativa de vida

	Estimativa	Erro Padrão	Valor t	Valor p
(Intercepto)	408.700	13.603	30.044	< 2e-16 ***
Log_PIB	-4.250	2.000	-2.125	0.0337 *
Escolaridade	-17.277	1.253	-13.784	< 2e-16 ***

Tabela 3: Modelo para Mortalidade Adulta

	Estimativa	Erro Padrão	Valor t	Valor p
(Intercepto)	-7.20709	0.37548	-19.195	< 2e-16 ***
Log_PIB	0.35068	0.05520	6.352	2.74e-10 ***
Escolaridade	0.75698	0.03459	21.881	< 2e-16 ***

Tabela 4: Modelo para Álcool

	Estimativa	Erro Padrão	Valor t	Valor p
(Intercepto)	-12.6810	1.9595	-6.472	1.28e-10 ***
Log_PIB	1.1985	0.2881	4.160	3.35e-05 ***
Escolaridade	3.4688	0.1805	19.213	< 2e-16 ***

Tabela 5: Modelo para IMC

A análise dos resultados dos modelos de regressão multivariada indica uma relação significativa entre as variáveis-preditoras, Log_PIB e Escolaridade, e as variáveis-resposta: Expectativa de Vida, Mortalidade Adulta, Consumo de Álcool e IMC.

No modelo para Expectativa de Vida, tanto Log_PIB quanto Escolaridade têm coeficientes positivos e altamente significativos (com valores p < 2e-16), sugerindo que um aumento em ambos os preditores está associado a uma maior expectativa de vida. O coeficiente da Escolaridade é particularmente elevado (2.00923), indicando que o nível educacional exerce um impacto substancial sobre a longevidade da população, mais do que o PIB.

Para Mortalidade Adulta, a relação se inverte, pois tanto Log_PIB quanto Escolaridade apresentam coeficientes negativos, indicando que um aumento nesses fatores está associado a uma redução na mortalidade adulta. Esse

resultado é coerente com a expectativa de que melhores condições econômicas e educacionais promovem a saúde e reduzem a mortalidade. A influência da Escolaridade é novamente mais intensa, com uma estimativa de -17.277, o que reflete a importância da educação na redução da mortalidade.

No modelo para Consumo de Álcool, os coeficientes de Log_PIB e Escolaridade são ambos positivos e significativos, sugerindo que esses fatores estão associados a um consumo maior de álcool. Esse resultado pode indicar que, em países com maior PIB e níveis educacionais mais altos, o consumo de álcool é mais elevado, possivelmente devido a padrões culturais ou de lazer associados a esses contextos.

Finalmente, para IMC, tanto Log_PIB quanto Escolaridade têm coeficientes positivos e significativos. Esse achado indica que países com PIB e escolaridade mais altos tendem a ter populações com IMC maior. Esse resultado pode estar relacionado a mudanças no estilo de vida, dieta e acesso a alimentos, que frequentemente acompanham o desenvolvimento econômico.

Esses modelos evidenciam que Escolaridade e Log_PIB impactam de forma consistente diferentes aspectos de saúde e bem-estar. A regressão multivariada permite captar essas relações complexas e interdependentes, ressaltando a importância de variáveis socioeconômicas para a saúde pública e qualidade de vida em diversas dimensões.

Após o ajuste dos modelos, avalia-se o efeito das variáveis preditoras sob o conjunto das variáveis dependentes, e não só individualmente. Para isso, utiliza-se o teste F da MANOVA, cujo resultado está disposto a seguir:

Fonte de Variação	Df	Pillai	F Aproximado	Num Df	Den Df	Pr(>F)
Log_GDP	1	0.55161	505.32	4	1643	< 2.2e-16 ***
Escolaridade	1	0.51730	440.19	4	1643	< 2.2e-16 ***
Resíduos	1646	-	-	-	-	-

Tabela 6: Resultados do MANOVA utilizando a estatística de Pillai

O teste indica que tanto Log_PIB quanto Escolaridade têm efeitos multivariados altamente significativos sobre o conjunto de variáveis-resposta (Expectativa de Vida, Mortalidade Adulta, Consumo de Álcool e IMC).

Para Log_PIB, a estatística de Pillai é 0.55161, com um valor de F aproximado de 505.32 e um valor-p extremamente pequeno. Esse resultado sugere que o Log_PIB tem um efeito significativo e substancial sobre as variáveis de saúde e bem-estar, indicando que o PIB impacta o bem-estar populacional em múltiplas dimensões.

No caso de Escolaridade, a estatística de Pillai é 0.51730, com um F aproximado de 440.19 e um valor-p também muito pequeno, confirmando que a escolaridade exerce uma influência significativa sobre o conjunto de variáveis-resposta. Esse resultado reforça a ideia de que a educação é um fator determinante para diferentes aspectos da saúde pública e da qualidade de vida.

A estatística de Pillai é particularmente útil em situações onde os efeitos multivariados podem ser sutis, pois tende a ser mais robusta em relação a desvios de normalidade e homogeneidade de covariâncias. Assim, esses resultados reforçam a importância das variáveis Log_PIB e Escolaridade como preditores significativos, justificando sua inclusão no modelo multivariado para analisar a saúde e o bem-estar em diversas dimensões.

3.3 Diagnóstico do modelo

Para verificar a qualidade e validade do modelo ajustado, é necessário testar as suposições exigidas, que são a normalidade, homocedasticidade e independência dos resíduos.

Para isso, foram feitos 4 tipos de gráficos para cada modelo, que podem ser vistos a seguir.

Para o primeiro modelo, o gráfico "Residuals vs Fitted"(Resíduo vs valor ajustado) indica uma distribuição relativamente aleatória dos pontos ao redor da linha horizontal, o que sugere que a relação entre as variáveis independentes e a dependente é aproximadamente linear. No entanto, há um leve padrão de dispersão dos resíduos que pode indicar uma pequena violação da homocedasticidade. No Q-Q Plot, observa-se uma leve curvatura nas caudas, sugerindo que os resíduos podem não ser perfeitamente normais. O Scale-Location indica que a variância dos resíduos parece ser aproximadamente constante ao longo dos valores ajustados. Isso apoia a suposição de homocedasticidade. Ou seja, Com base nesses gráficos, as suposições de linearidade, homocedasticidade e normalidade dos resíduos são, em geral, atendidas de forma satisfatória, com apenas algumas pequenas violações que provavelmente não comprometem significativamente o modelo. Observações com alta alavancagem podem merecer uma análise mais aprofundada para garantir que não estejam distorcendo os resultados.

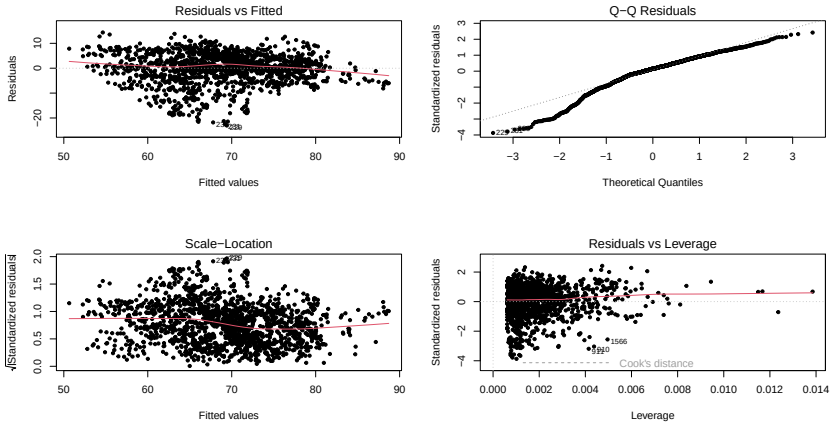

Figura 3: Resíduos do Modelo para Expectativa de vida

Para o modelo de Mortalidade Adulta, os gráficos de diagnóstico indicam algumas potenciais violações das suposições do modelo de regressão. No gráfico de Resíduos vs. Valores Ajustados, os resíduos não parecem distribuídos de forma totalmente aleatória ao redor da linha zero, sugerindo uma leve heterocedasticidade, o que poderia impactar a robustez das estimativas. No gráfico Q-Q, observa-se um desvio considerável dos pontos em relação à linha reta, principalmente nas caudas, o que sugere que os resíduos não seguem uma distribuição normal. Essa falta de normalidade dos resíduos pode comprometer a validade das inferências estatísticas. O gráfico Scale-Location também indica possível heterocedasticidade, pois a dispersão dos resíduos não se mantém constante ao longo dos valores ajustados. Finalmente, o gráfico de Resíduos vs. Alavancagem revela pontos com alta alavancagem e alta distância de Cook, sugerindo que alguns pontos influentes podem estar distorcendo o modelo.

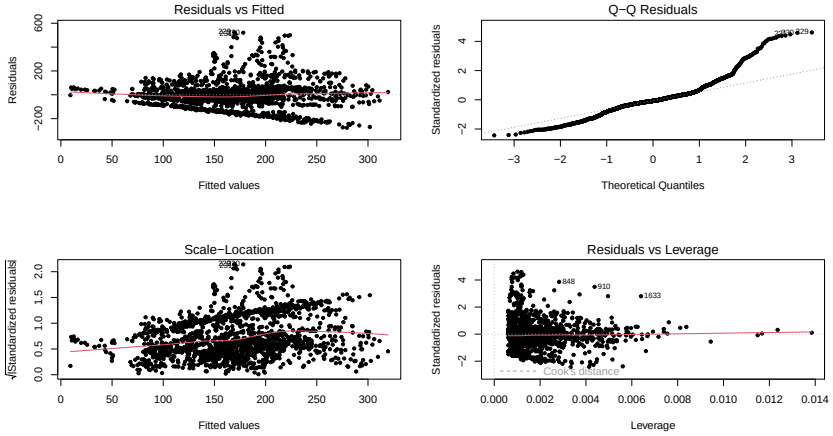

Figura 4: Resíduos do Modelo para Mortalidade Adulta

Para o modelo de Álcool, os gráficos de diagnóstico também apresentam alguns problemas semelhantes, embora em menor intensidade em relação ao modelo de Mortalidade Adulta. No gráfico de Resíduos vs. Valores Ajustados, os resíduos estão distribuídos de forma ligeiramente não uniforme ao redor da linha zero, o que pode indicar uma leve heterocedasticidade. O gráfico Q-Q também mostra um desvio em relação à linha reta, especialmente nas extremidades, o que sugere que a normalidade dos resíduos não é totalmente atendida. No gráfico Scale-Location, a dispersão dos resíduos não se mantém uniforme ao longo dos valores ajustados, reforçando a possibilidade de heterocedasticidade.

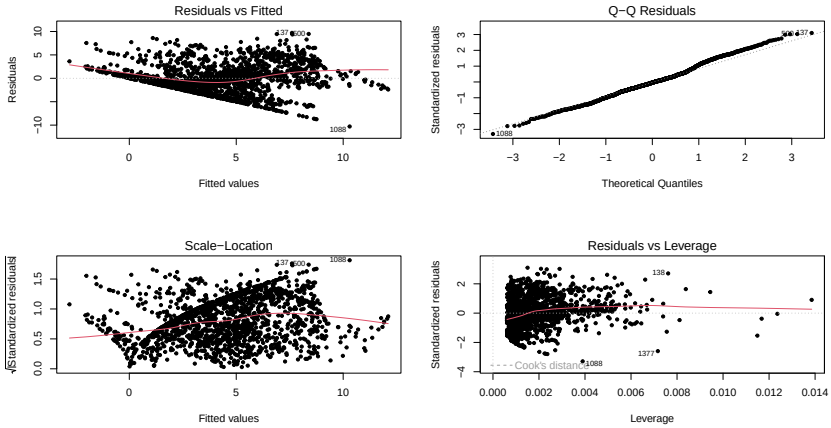

Figura 5: Resíduos do Modelo para Álcool

Assim como os outros modelos, os resíduos do modelo de IMC, revelam alguns problemas importantes com as suposições do modelo de regressão.

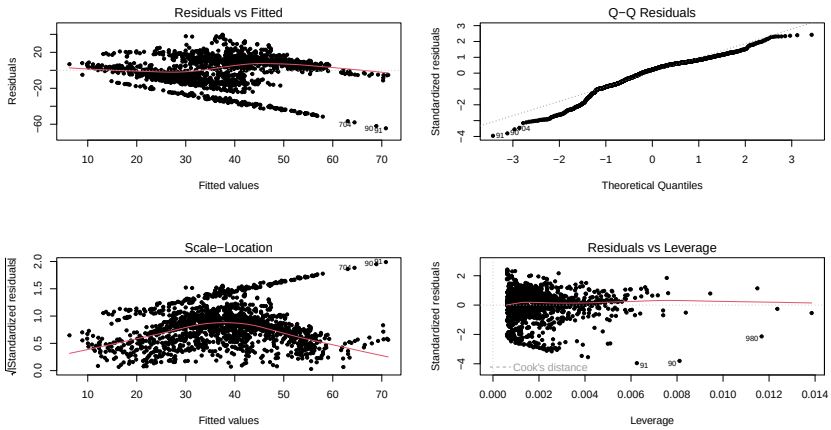

Figura 6: Resíduos do Modelo para IMC

O Teste de Mardia, cujo resultado está demonstrado na próxima tabela, será usado para testar a normalidade multivariada dos resíduos. O resultado indica assimetria e violação da curtose, e que, portanto, os resíduos do modelo não seguem uma distribuição normal multivariada.

Teste	Estatística	Valor-p	Resultado
Mardia Skewness	2112.41	0	Não
Mardia Kurtosis	27.32	0	Não
MVN	-	-	Não

Tabela 7: Resultados do teste de normalidade multivariada de Mardia

Assim, após a analisar os resíduos, observou-se que algumas suposições da regressão linear não foram completamente atendidas, como a normalidade dos resíduos e a homocedasticidade. Tentativas de transformação dos dados foram realizadas para melhorar o ajuste e o cumprimento dessas suposições, mas essas modificações resultaram em perda de significância estatística nos modelos.

Dessa forma, optou-se por manter os modelos na forma original, pois, apesar das limitações, eles produzem resultados significativos e fornecem interpretações úteis. Embora seja ideal que as suposições da regressão linear sejam plenamente atendidas, é possível prosseguir com a análise, desde que se reconheça e se discuta as limitações encontradas. Assim, os resultados obtidos ainda podem oferecer insights valiosos sobre as relações entre as

variáveis, respeitando as ressalvas sobre a confiabilidade das inferências.

4 Vantagens, desvantagens e limitações

O modelo de regressão multivariada é uma ferramenta estatística poderosa que permite investigar como múltiplas variáveis preditoras influenciam simultaneamente várias variáveis-resposta. Esse tipo de análise oferece vantagens importantes em cenários onde as respostas são inter-relacionadas, pois permite modelar a correlação entre as respostas, o que pode aumentar a precisão das estimativas e permitir uma compreensão mais completa das relações entre as variáveis. No contexto de um estudo como este, em que variáveis como expectativa de vida, mortalidade adulta, consumo de álcool e IMC podem estar associadas a fatores econômicos e educacionais, a regressão multivariada possibilita uma análise integrada e mais robusta dos efeitos conjuntos do PIB e da escolaridade.

Entre as vantagens da regressão multivariada, destaca-se a capacidade de considerar e modelar a interdependência entre as respostas, o que melhora a eficiência estatística em relação a análises separadas univariadas. Além disso, ao utilizar um modelo único, evita-se o problema de múltiplas comparações que surgiria ao realizar regressões independentes para cada resposta. Esse método também facilita a interpretação dos efeitos dos preditores ao oferecer uma visão holística sobre como variáveis de interesse impactam simultaneamente múltiplos desfechos.

Por outro lado, o modelo de regressão multivariada apresenta algumas desvantagens e limitações. Em primeiro lugar, ele depende de suposições estatísticas rigorosas, como a normalidade multivariada dos resíduos, a linearidade entre as variáveis e a homogeneidade das variâncias-covariâncias. No presente estudo, as análises de resíduos mostraram que essas suposições não foram totalmente atendidas, com evidências de falta de normalidade nos resíduos e problemas de ajuste em alguns gráficos. Embora esses desvios das suposições não impeçam necessariamente o uso do modelo, eles indicam que as inferências podem não ser totalmente confiáveis. As tentativas de aplicar transformações para melhorar o ajuste do modelo não foram bem-sucedidas, pois o modelo deixou de ser significativo após as transformações. Essa situação destaca uma limitação prática importante: muitas vezes, transformações não resolvem problemas de aderência aos pressupostos e podem até enfraquecer a utilidade do modelo.

Outro ponto a considerar é que, quando os dados apresentam uma estrutura complexa ou desvio das suposições, o modelo de regressão multivariada pode ser sensível a outliers e influências excessivas, como foi evidenciado pelas análises de alavancagem e distância de Cook realizadas. Nesses casos, a interpretação dos coeficientes e dos efeitos pode ser distorcida, e ajustes ou técnicas alternativas podem ser necessários.

Por fim, apesar das limitações observadas, o modelo de regressão multivariada ainda se mostra uma ferramenta útil para o contexto do estudo, pois possibilita uma análise conjunta das variáveis e permite extrair insights significativos. A decisão de prosseguir com o modelo sem transformações foi baseada na perda de significância observada ao tentar ajustá-lo aos pressupostos, e os resultados ainda fornecem informações valiosas sobre o impacto do PIB e da escolaridade nas variáveis de saúde. Contudo, é importante interpretar os achados com cautela e considerar as limitações ao discutir as implicações dos resultados.

5 Conclusão

A análise do presente estudo visou investigar as relações entre fatores socioeconômicos, como PIB per capita e escolaridade, e variáveis de saúde pública, incluindo expectativa de vida, mortalidade adulta, consumo de álcool e IMC, utilizando um modelo de regressão multivariada. A partir de um conjunto de dados extenso, foi realizada uma análise descritiva e exploratória inicial, que revelou padrões importantes, como a associação positiva entre escolaridade e expectativa de vida e a associação negativa entre PIB e mortalidade adulta.

Na etapa de modelagem, optou-se pela regressão multivariada, considerando a interdependência das variáveis de saúde, o que permitiu uma análise integrada e mais robusta dos efeitos dos fatores preditores. Durante o diagnóstico do modelo, observou-se que algumas das suposições estatísticas, como a normalidade multivariada dos resíduos, não foram plenamente atendidas. Tentativas de transformação para ajustar o modelo às suposições não foram eficazes, pois resultaram em perda de significância. Apesar dessas limitações, os resultados fornecem uma compreensão abrangente das relações entre as variáveis estudadas e ressaltam a relevância do PIB e da escolaridade para as condições de saúde pública.

Assim, conclui-se que, embora o modelo apresente algumas limitações em relação ao ajuste aos pressupostos, ele ainda oferece insights valiosos. A interpretação dos resultados deve ser feita com cautela, levando em conta as limitações apontadas, mas os achados sugerem que investimentos em educação e melhorias econômicas podem ter impactos positivos nas condições de saúde da população.

6 Referências

- Richard A. Johnson & Dean W. Wichern. *Applied Multivariate Statistical Analysis, Sexta edição*.
- Mardia, K. V., Kent, J. T., & Bibby, J, M. (1979). Multivariate Analysis. Academic Press.
- Getting started with Multivariate Multiple Regression - UVA Library