Laboratório 3 - Introdução à Regressão Linear Simples

Camille Menezes e Michel Miler

15 de abril de 2023

1 Introdução

Este relatório tem como objetivo analisar a relação entre as taxas de desemprego e os índices de suicídio nos Estados Unidos, utilizando dados do arquivo "`desemprego.csv`"que abrange o período de 1950 a 2019. O índice de suicídio é apresentado por cada 1000 habitantes.

Para alcançar este objetivo, serão utilizadas técnicas de regressão linear para verificar se há uma relação linear entre as variáveis em questão. Serão estimadas as variâncias dos coeficientes de regressão e testada a significância do modelo. Adicionalmente, um intervalo de confiança será obtido para os parâmetros do modelo com um nível de confiança de 95%.

Além das análises de regressão linear e de teste de significância do modelo, será efetuada uma análise de resíduos para verificar se as hipóteses de normalidade, homogeneidade de variâncias e independência dos erros estão sendo satisfeitas pelo modelo. Para isso, serão utilizados gráficos de probabilidade normal dos resíduos e teste de normalidade de *Shapiro-Wilk*, gráficos de dispersão dos resíduos em relação aos valores ajustados e teste de *Breush-Pagan* para verificar homocedasticidade, e o teste de *Durbin-Watson* para verificar a independência dos erros.

Todas as análises necessárias serão realizadas com o auxílio do software R (R Core Team, 2022). Com base nos resultados obtidos, serão tiradas conclusões sobre a capacidade do modelo em explicar a relação entre o desemprego e o índice de suicídios nos EUA.

Por fim, é importante destacar que o banco de dados utilizado contém informações sobre o desemprego e a taxa de suicídio em determinados anos. Cada linha representa um ano específico, enquanto as colunas representam o desemprego e a taxa de suicídio, respectivamente.

2 Resultados

2.1 Análise descritiva

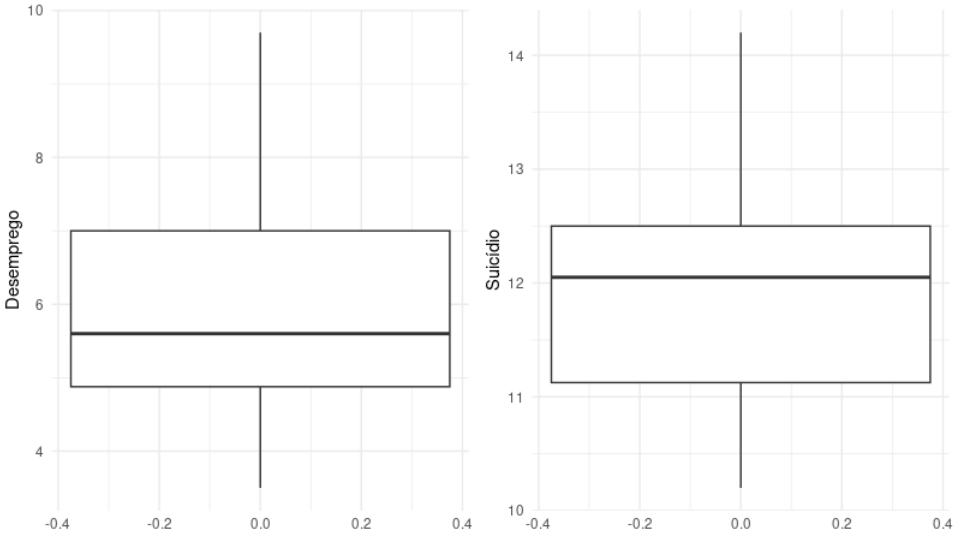
Através da Tabela 1, é possível notar que taxa de desemprego variou entre 3,5% e 9,7% no período de 1950 a 2019, enquanto o índice de suicídios variou entre 10,2 e 14,2. A média, mediana e moda para ambas as variáveis estão próximas, indicando uma distribuição relativamente simétrica dos dados. A variância da taxa de desemprego é maior do que a do índice de suicídios, o que sugere que a taxa de desemprego apresentou maior variabilidade no período analisado.

Tabela 1: Sumarização das variáveis `taxa de desemprego` e `índice de suicídios` dos EUA no período de 1950 a 2019

	Desemprego	Suicídio
Mínimo	3,50	10,20
1º quartil	4,88	11,12
Mediana	5,60	12,05
Média	5,60	11,98
Moda	5,60	12,50
3º quartil	7,00	12,50
Máximo	9,70	14,20
Variância	2,60	0,910

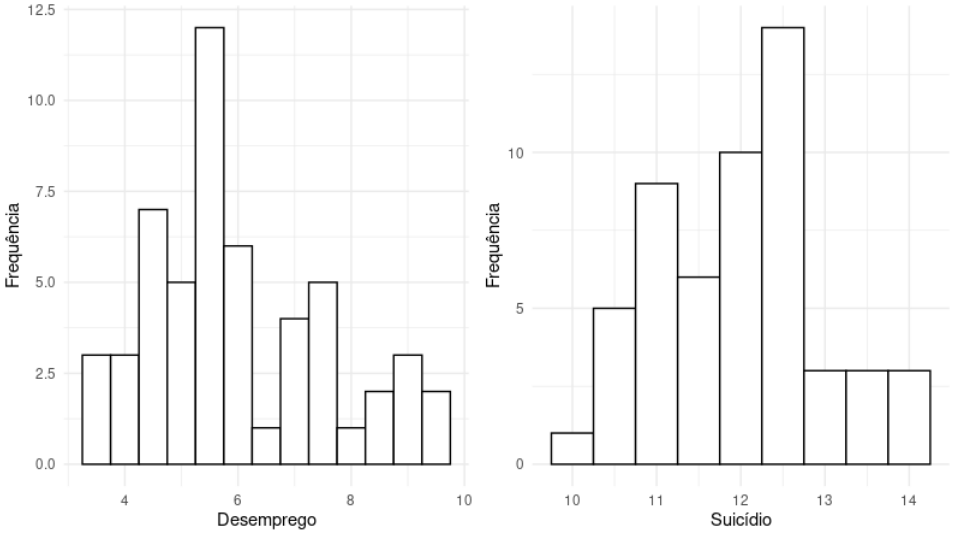
Pelos boxplots da Figura 1, é percebido que a taxa de desemprego apresentou uma maior variabilidade entre a mediana e o terceiro quartil, enquanto a taxa de suicídio apresentou uma maior variabilidade entre o primeiro quartil e a mediana. Essas características podem indicar não simetria, contrapondo-se com o que foi avaliado anteriormente. Contudo, nesse primeiro momento, não foi possível notar a presença de *outliers*.

Figura 1: *Boxplots* da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019



Os histogramas da Figura 2, mostram que a distribuição da taxa de desemprego e da taxa de suicídio tende a unimodalidade, o que é uma das características da distribuição normal.

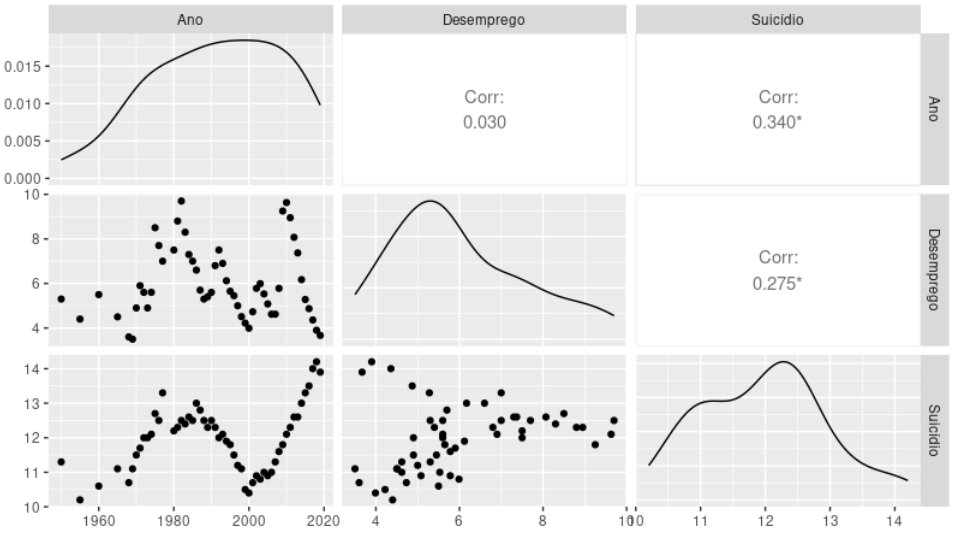
Figura 2: Histogramas da taxa de desemprego e do índice de suicídios dos EUA no período de 1950 a 2019



Na Figura 3, é possível verificar um gráfico de dispersão que apresenta a relação entre a taxa de desemprego e a taxa de suicídio. Embora pareça haver uma relação de linearidade entre as duas variáveis, é importante destacar que a correlação entre elas é fraca, conforme observado pelo coeficiente de correlação e pelos pontos que fogem completamente da linha de tendência.

Vale ressaltar que cada observação representa a taxa de desemprego e suicídio em um determinado ano, indicando uma relação temporal entre elas e, portanto, falta de independência entre as observações. Essa relação temporal pode ser percebida também no gráfico de dispersão entre o ano e a taxa de suicídio, no qual é possível identificar tendências em relação à taxa de suicídio em cada ano, sendo que a partir dos anos 2000 há um crescimento linear bastante expressivo na taxa de suicídio. Excelente!

Figura 3: Correlação e gráficos de dispersão e densidade da taxa de desemprego e do índice de suicídios dos EUA no período de 1950 a 2019



2.2 Modelo linear de regressão

2.2.1 Modelo I

A Tabela 2, apresenta uma sumarização da regressão linear em que o índice de desemprego é utilizado para predizer o suicídio. A estimativa do intercepto indica que quando o índice de desemprego é igual a zero, espera-se um número médio de suicídios de 11,00. A estimativa do coeficiente angular aponta que para cada aumento de uma unidade no índice de desemprego, espera-se um aumento médio de 0,16 na taxa de suicídios.

O modelo de regressão linear sugere que o índice de desemprego pode ser um fator significativo na predição do número de suicídios nos Estados Unidos. Uma vez que, o valor-p para o coeficientes de interceptação e angular indicam o que esses coeficientes são estatisticamente significativos a um nível de significância de 5%

Os intervalos de confiança (IC) para os coeficientes também são apresentados na tabela. Indicando que, com uma confiança de 95%, o intervalo conterá o verdadeiro valor do coeficiente.

Tabela 2: Sumarização da regressão linear em que o índice de desemprego é utilizado para predizer o suicídio nos EUA no período de 1950 a 2019

Coeficientes	Estimativa	EP	VAR	Est. de teste	P-valor	IC (2,5%)	IC (97,5%)
β_0	11,00	0,49	0,240	22,46	0,00	10,02	11,98
β_1	0,16	0,08	0,006	2,06	0,04	0,004	0,32

A estatística F é significativa na Tabela 3, o valor-p sugere que o modelo é significativo, ou seja, há alguma relação entre a taxa de desemprego e o índice de suicídio.

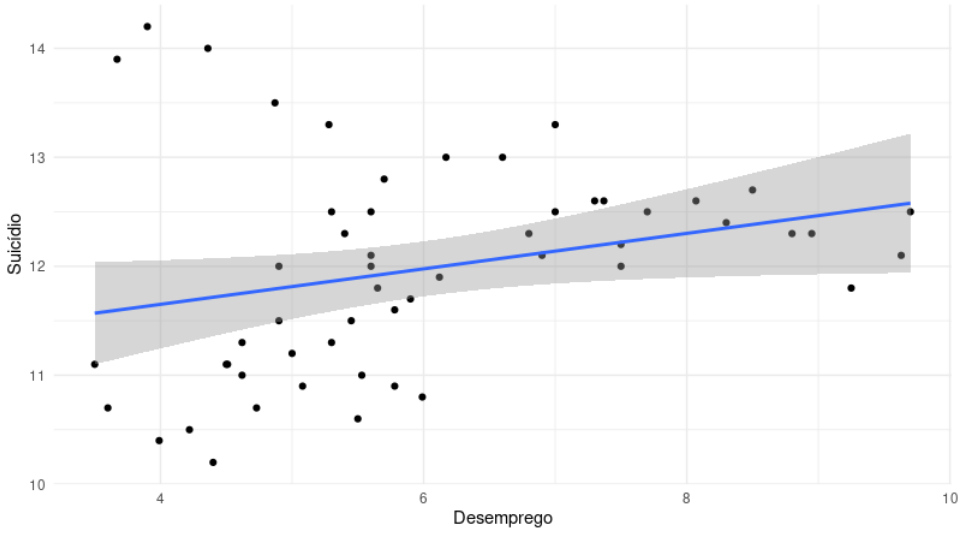
Apesar dos coeficientes serem significativos, a tabela evidencia ainda um ajuste fraco do modelo aos dados. O coeficiente de determinação (R^2), que é ligeiramente menor que o R^2_a ajustado, indica que apenas 8% da variância no índice de suicídio pode ser explicada pelo modelo que, para esse caso de modelo linear simples, é apenas a variável taxa de desemprego.

Tabela 3: Coeficiente de determinação e teste F da regressão linear da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019

R^2	R^2_a	Estatística	P-valor
0,08	0,06	4,26	0,04

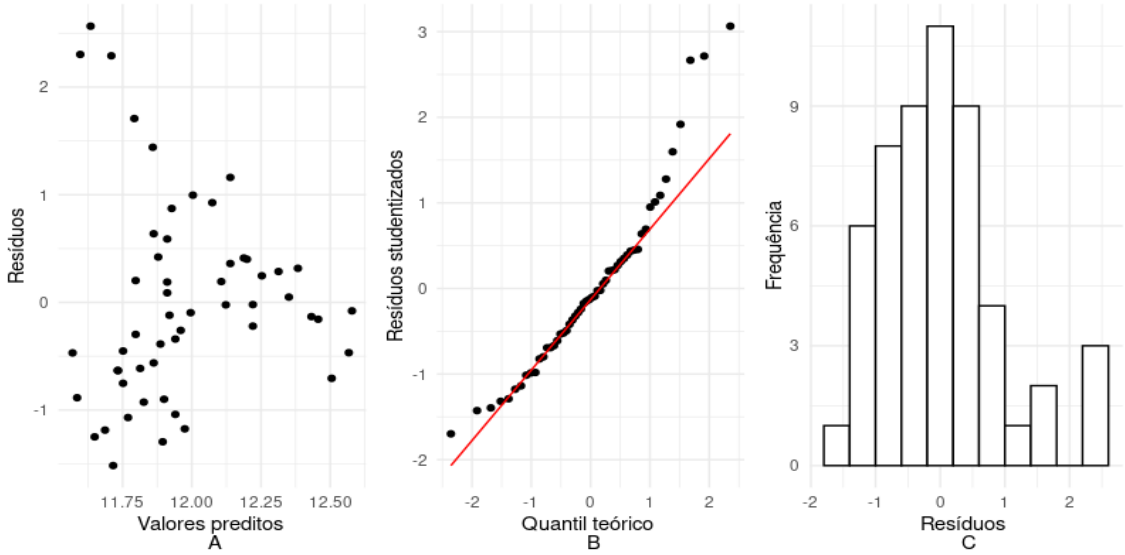
Olhando para o gráfico de dispersão da Figura 4, podemos observar que grande parte dos pontos fogem do intervalo de confiança para o valor esperado da taxa de suicídio dado a taxa desemprego estimada pelo modelo. O modelo explica de maneira bastante pobre a taxa de suicídio.

Figura 4: Gráfico de dispersão e curva de regressão linear ajustada aos dados da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019



O gráfico de resíduos contra valores preditos na Figura 5 (A), mostra que parece haver uma maior variabilidade nos resíduos entre os menores valores preditos, ou seja, a variação dos resíduos está diminuindo a medida que os valores preditos crescem. Realizando o teste de *Breusch-Pagan*, foi obtido um p-valor de aproximadamente zero, rejeitando a hipótese nula, ao nível de significância de 5%, de que a variância dos resíduos é constante. Portanto, os resíduos violam a suposição de normalidade.

Figura 5: Gráfico dos resíduos versus valores ajustados, resíduos studentizados versus quantil teórico e histograma dos resíduos da regressão linear da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019



O *qq-plot* dos resíduos studentizados, Figura 5 (B), mostra que a normalidade dos resíduos está sendo prejudicada pelas caudas da distribuição, principalmente na cauda direita, onde os resíduos estão sendo superestimados pelo modelo. O histograma dos resíduos, Figura 5 (C), reforça essa ideia, embora a variável aparenta ter uma distribuição de simétrica, há valores discrepantes acima do valor de resíduo igual a 2. O teste de *Shapiro-Wilk* indicou um p-valor de aproximadamente zero, rejeitando a hipótese nula, a um nível de significância

de 5%, de que os resíduos vem de uma distribuição normal. Portanto, com base na análise gráfica e no teste, podemos afirmar que a suposição de normalidade dos erros foi violada pelo modelo.

Com base nessa análise, optamos por retirar as observações que estão com resíduos com 2 desvios padrões acima da média, considerando-as como "atípicas", ajustando o modelo de regressão novamente. Entretanto, essa decisão talvez não seja a mais adequada, já que essas observações foram sequenciais e dos anos mais recentes (2017, 2018 e 2019), os valores da taxa de suicídio delas não são discrepantes das demais (razão pela qual o *boxplot* não as indicou como *outliers*). Essas observações podem estar sendo problemáticas devido à presença de uma ou mais variáveis "ocultas"que deveriam estar no modelo, que exigiriam um modelo de regressão linear múltiplo ou outras técnicas estatísticas que incorporassem os anos.

2.2.2 Modelo II

A Tabela 4 apresenta os coeficientes estimados do modelo, enquanto a Tabela 5 apresenta o coeficiente de determinação R^2 e o teste F.

Observa-se que o intercepto indica que quando o índice de desemprego é igual a zero, o número de suicídios esperado é de aproximadamente 10,11. Já o coeficiente angular aponta que o aumento de uma unidade no índice de desemprego está associado a um aumento esperado de 0,29 no número de suicídios. Ambos os coeficientes são estatisticamente significativos, com p-valor aproximadamente zero para ambos.

O coeficiente de determinação apresentado indica que o modelo é capaz de explicar apenas 29% da variabilidade dos dados, o que é razoavelmente maior do que foi obtido pelo primeiro modelo ajustado, entretanto, ainda é um valor pequeno de explicabilidade. O teste F sugere que o modelo é estatisticamente significativo, já que o p-valor foi aproximadamente zero.

Tabela 4: Sumarização da regressão linear em que o índice de desemprego é utilizado para predizer o suicídio nos EUA no período de 1950 a 2019

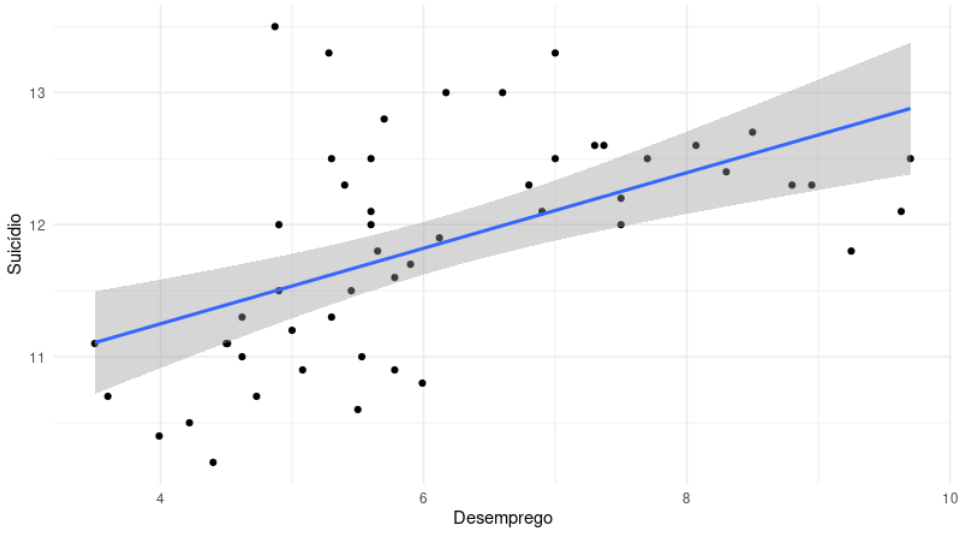
Coeficientes	Estimativa	EP	VAR	Est. de teste	P-valor	IC (2,5%)	IC (97,5%)
β_0	10,11	0,40	0,160	25,23	0,00	9,30	10,91
β_1	0,29	0,06	0,004	4,51	0,00	0,16	0,41

Tabela 5: Coeficiente de determinação e teste F da regressão linear da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019

R^2	R^2_a	Estatística	P-valor
0,29	0,28	20.33	0,00

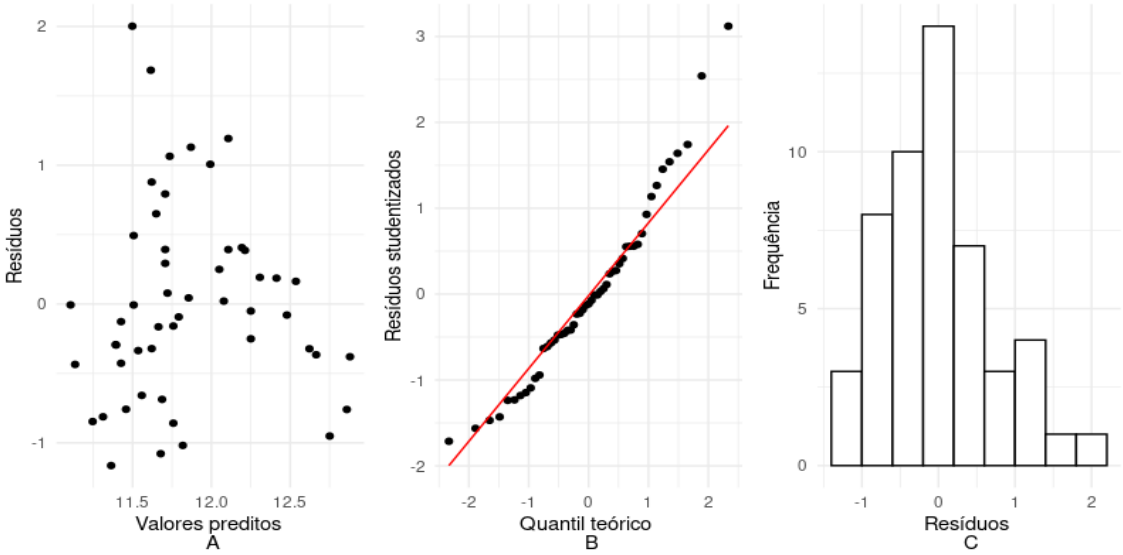
As observações ainda continuam extrapolando o intervalo de confiança do valor esperado da taxa de suicídio dado a taxa de desemprego estimado pelo modelo, como é possível observar na Figura 6.

Figura 6: Gráfico de dispersão e curva de regressão linear ajustada aos dados da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019



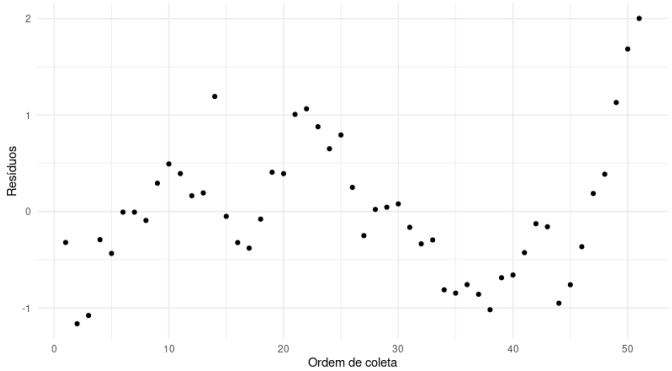
No gráfico de resíduos versus valores preditos na Figura 7 (A), é possível observar que a variabilidade dos resíduos não parece ser constante, apesar de não apresentar o mesmo padrão do modelo anterior. Para maiores e menores valores preditos, os resíduos tendem a ser menores do que zero. Entretanto, realizando o teste de *Breusch-Pagan*, o p-valor obtido foi aproximadamente 0,2, indicando a não rejeição da hipótese nula (homocedasticidade) ao nível de significância de 5%. Portanto, tendo em vista a análise gráfica e o teste, não é possível afirmar que a suposição de variância constante dos resíduos foi violada.

Figura 7: Gráfico dos resíduos versus valores ajustados, resíduos studentizados versus quantil teórico e histograma dos resíduos da regressão linear da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019



O *qq-plot* dos resíduos apresenta menos problemas nas caudas do que no modelo anterior, entretanto ainda há resíduos superestimados pelo modelo que podem afetar a suposição de normalidade, Figura 7 (B). O histograma dos resíduos aparenta seguir uma distribuição normal padrão, entretanto é possível ver uma leve assimetria à direita devido ao problema da cauda à direita, reforçando o que foi observado no *qq-plot*. O teste de Shapiro Wilk resultou em um p-valor de aproximadamente 0,1, o que indica a não rejeição da hipótese nula de que os resíduos vieram de uma distribuição normal ao nível de significância de 5%. Portanto, apesar dos problemas citados, podemos dizer que a suposição de normalidade dos resíduos não foi violada.

Figura 8: Gráfico dos resíduos versus ordem de coleta da regressão linear da taxa de desemprego e índice de suicídios dos EUA no período de 1950 a 2019



O gráfico dos resíduos versus a ordem de coleta apresenta as mesmas tendências do gráfico de dispersão do ano versus a taxa de suicídio, o que reforça o que foi dito anteriormente sobre a não independência das observações. Realizando o teste de *Durbin-Watson* para verificar a independência dos resíduos, foi obtido um p-valor de aproximadamente zero, rejeitando a hipótese nula de que não há correlação serial dos resíduos, ou seja, rejeitando a hipótese de independência dos resíduos.

3 Conclusão

Parabéns pela conclusão apresentada. Vocês realizaram uma análise bem detalhada e conseguiram sumarizar muito bem aqui.

Em conclusão, este relatório apresenta uma análise da relação entre as taxas de desemprego e os índices de suicídio nos Estados Unidos, utilizando dados do período de 1950 a 2019. Uma análise descritiva foi realizada e dois modelos de regressão linear foram ajustados aos dados.

O primeiro modelo apresentou coeficientes significativos, mas um coeficiente de determinação baixo, sugerindo que a taxa de desemprego pode não ser um bom preditor do índice de suicídios. Por outro lado, o segundo modelo apresenta coeficientes significativos e um coeficiente de determinação mais alto, indicando um ajuste superior. Além disso, o teste F para o segundo modelo foi mais significativo do que para o primeiro, sugerindo uma relação mais forte entre a taxa de desemprego e o índice de suicídios.

No entanto, o primeiro modelo violou todas as suposições necessárias para os resíduos, mostrando que é inadequado para explicar a taxa de suicídio. O segundo modelo, com as observações consideradas atípicas retiradas, não violou a suposição de normalidade e homoscedasticidade, mas violou a suposição de independência dos resíduos. Apesar de melhor ajustado e mais significativo que o primeiro, o segundo modelo tende a ser um mau previsor da taxa de suicídio para os próximos anos, pois as observações dos últimos três anos foram retiradas.

Pode-se concluir que, embora haja uma relação linear entre as taxas de desemprego e os índices de suicídio nos Estados Unidos, a relação não é tão clara, as suposições do modelos foram violadas e outras variáveis podem estar influenciando o resultado. Logo, nenhum dos dois modelos é adequado para modelar esses dados.

É importante salientar ainda que o banco de dados utilizado apresentou algumas limitações, como a falta de outras variáveis que podem estar correlacionadas com a taxa de suicídio, como a disponibilidade de serviços de saúde mental. Recomenda-se, portanto, a realização de novos estudos para aprofundar a análise da relação entre as variáveis.

Referências

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