LABORATÓRIO 3: Regressão Linear - Desemprego nos EUA entre 1950 a 2019

Fernando Bispo, Jeff Caponero

Sumário

Introdução

Composto por 54 observações, o conjunto de dados em estudo traz informações pertinentes o Índice de desemprego e a Índice de suicídio nos EUA no período entre 1950 e 2019.

O presente relatório tem como objetivo a introdução das técnicas de Regressão Linear Simples e a pratica da elaboração de relatórios analíticos fundamentados na Análise Exploratória de Dados, preenchendo assim os pré-requisitos solicitados para o conjunto de dados proposto.

Metodologia

Serão realizadas análises sobre o levantamento das Índices de desemprego e o índice de suicídios nos EUA para o período de 1950 a 2019. Ressalta-se que o índice de suicídios foi calculado para cada 1000 habitantes.

As variáveis contidas no arquivo "desemprego.csv" são:

- Ano (**ano**);
- Índice de Desemprego por 1000 habitantes (**desemp**);
- Índice de Suicídio por 1000 habitantes (**suic**).

Objetivos

O objetivo dessa análise visa responder aos seguintes tópicos:

a. Identificar, por meio da análise dos dados, se a Índice de suicídios é função linear do desemprego.
b. Obter as estimativas das variâncias de β_0 e β_1.
c. Testar a significância do modelo e reportar a conclusão obtida a um nível de significância de 5%
d. Obter os intervalos de confianças para os parâmetros do modelo com o nível de 95% de confiança e interpretar os resultados.

Análise dos dados

Análise Preliminar

A Tabela 1 traz as principais medidas resumo das variáveis em análise, viabilizando assim uma análise preliminar desses dados.

Tabela 1: Medidas resumo dos Índices de desemprego e suicídio nos EUA de 1950 a 2019

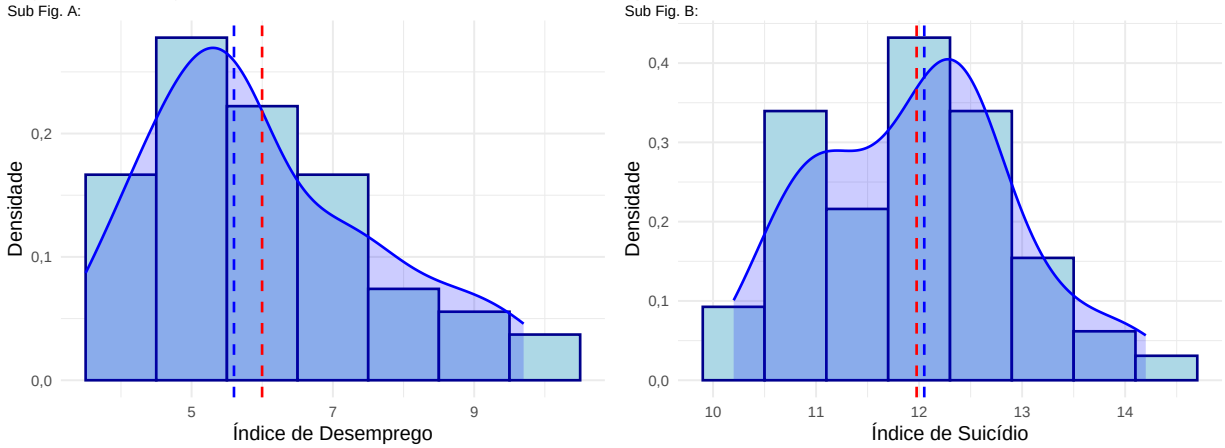
	DESEMP	SUIC
Mín	3,50	10,20
Q1	4,87	11,10
Med	5,60	12,05
Média	6,00	11,98
Q3	7,00	12,50
Máx	9,70	14,20
Desv.padrão	1,61	0,95
CV	0,27	0,08
Assimetria	0,63	0,22
Curtose	-0,50	-0,56

Legenda:
[1] DESEMP = Índice de Desemprego
[2] SUIC = Índice de Suicídio
Fonte: Laboratório 3, Disciplina Análise de Regressão 2023.1.

Para uma primeira análise é possível concluir que a variável Índice de Desemprego apresenta uma maior variabilidade dos dados, em comparação com o Índice de Suicídio, fato esse constatado pelo Coeficiente de Variação, caracterizando assim uma maior homogeneidade dos dados obtidos referente ao o Índice de Suicídio.

Com relação ao Coeficiente de Assimetria, ambas as variáveis possuem valores de assimetria positiva, indicando que a maioria dos valores são menores que a média. Já com base no Coeficiente de Curtose é possível identificar um comportamento Platocúrtico dos dados, ou seja, um comportamento mais achatado da distribuição dos dados. A análise gráfica facilitará a identificação das informações trazidas pela tabela em análise.

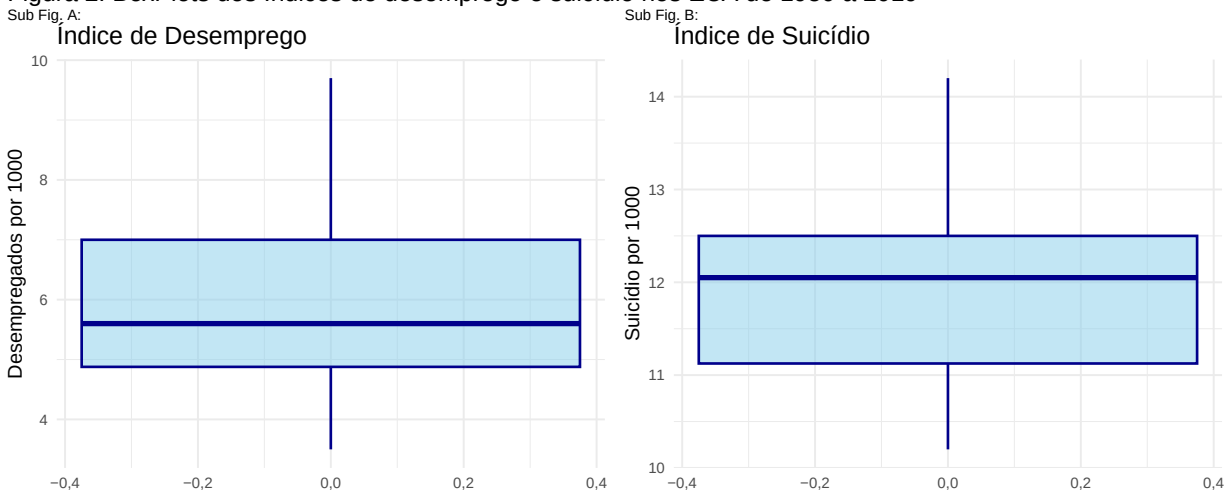
Figura 1: Histograma e Densidade dos Índices de desemprego e suicídio nos EUA de 1950 a 2019



Nota: Média representada pela linha vertical tracejada vermelha.
Mediana representada pela linha vertical tracejada azul.

Conforme identificado na Tabela 1, a Figura 1 traz a representação gráfica das conclusões realizadas para os dados, no que diz respeito a assimetria e curtose.

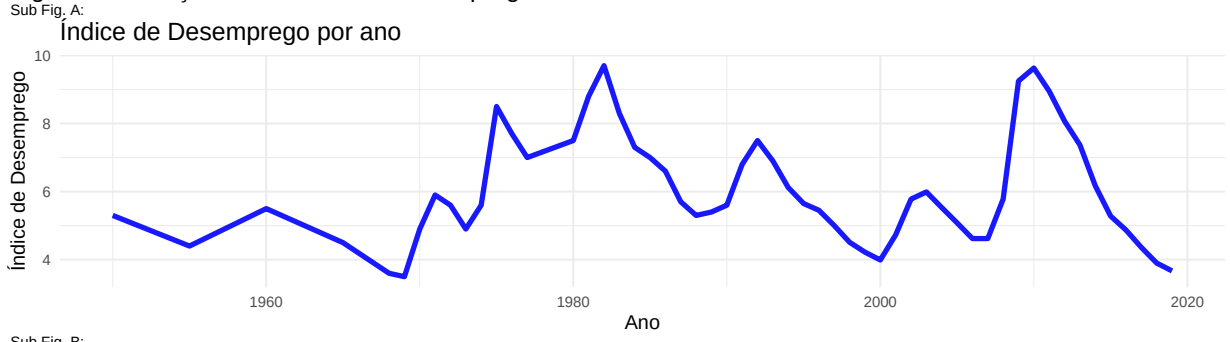
Figura 2: BoxPlots dos Índices de desemprego e suicídio nos EUA de 1950 a 2019



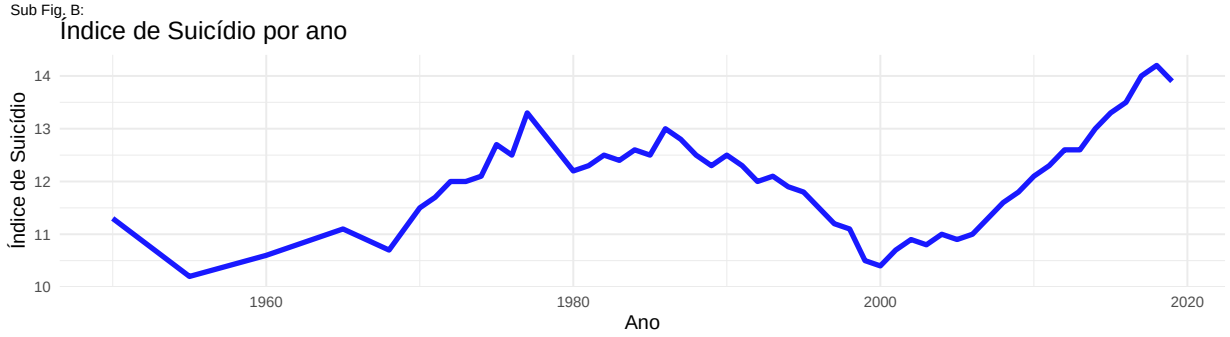
A análise dos *BoxPlots*, Figura 2, traz pouco mais informação, uma vez que agora é possível verificar com mais facilidade a inexistência de valores discrepantes (*outliers*) que poderiam influir negativamente na regressão linear.

Figura 3: Evolução dos Índices de Desemprego e Suicídio nos EUA entre 1950 e 2019

Sub Fig. A:



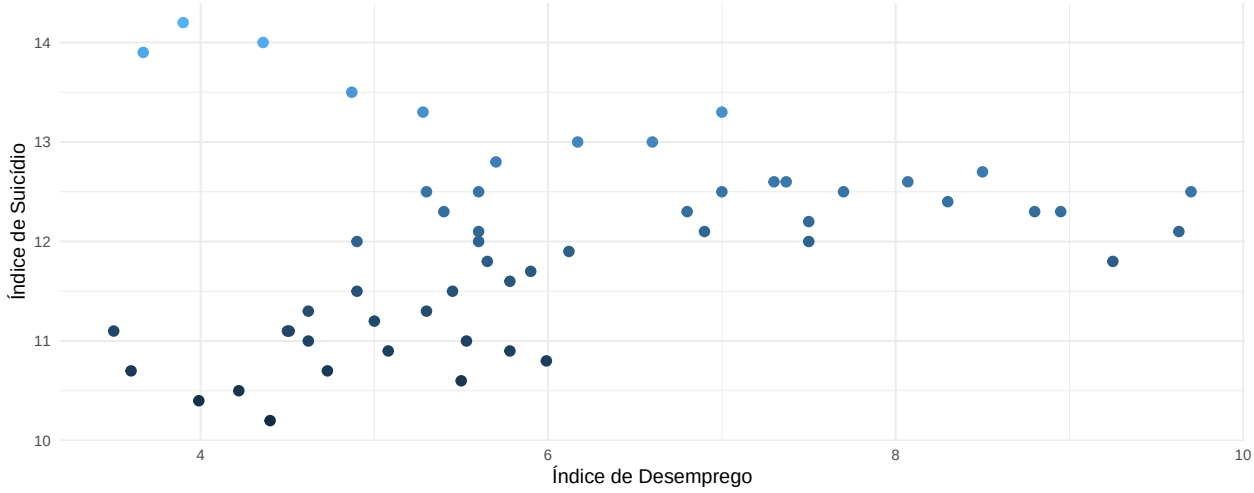
Sub Fig. B:



Através da Figura 3 é possível avaliar a evolução dos índices em analise ao longo do tempo, em que é possível identificar uma queda no Índice de Desemprego nos anos finais em que os dados foram coletados, contudo o Índice de Suicídio está apresentando um sinal de queda após um longo período de alta.

A Figura 4 foi desenvolvida a fim de identificar a relação entre as variáveis em análise.

Figura 4: Relação entre Índice de Desemprego e a Índice de Suicídio entre 1950 e 2019, nos EUA



Após análise da Figura 4 não se identificou uma evidente correlação entre as variáveis em análise. Para avaliar a força dessa possível correlação foi realizado o cálculo do Coeficiente de Correlação de Pearson ($\hat{\rho}$), medida que avalia o grau da correlação linear entre variáveis, em que se obteve o valor de 0.2752, caracterizando uma baixa relação entre as variáveis.

Para avaliar a significância da correlação segue a Tabela 2 com os resultados do Teste de Hipóteses com nível de significância de 5% que tem como hipóteses:

$$H_0 : \hat{\rho} = 0$$
$$H_1 : \hat{\rho} \neq 0.$$

Tabela 2: Teste de Hipótese para Correlação

	Resultados
t	2,06387
p-valor	0,04404
LI	0,00799
LS	0,50566

Legenda:
[1] t: Estatística t-Student
[2] p-valor: Menor nível de significância para o qual rejeita-se H0 com os dados observados
[3] LI: Limite Inferior do Intervalo de Confiança
[4] LS: Limite Superior do Intervalo de Confiança
Nota: Teste realizado com 5% de significância

Conforme expresso na Tabela 2, levando em consideração o **p-valor** a Hipótese Nula foi rejeitada, e com 95% de confiança se pode afirmar que **é significativa a relação linear entre as variáveis em estudo.**

Regressão Linear

Tendo em vista a existência de correlação linear entre as variáveis, foi realizado o ajuste do modelo cuja equação é apresentada a seguir:

$$\hat{Y}_i = 10,9994 + 0,1628X_i, i = 1, ..., n.$$

Significância do modelo

Tendo em vista a necessidade de se avaliar a significância dos parâmetros, o teste de hipótese para tal situação será realizado, contendo as seguintes hipóteses:

$$H_0 : \hat{\beta}_0 = 0$$
$$H_1 : \hat{\beta}_0 \neq 0.$$

As Tabelas 3, 4 e 5 trazem os principais resultados da sumarização dos dados do modelo ajustado, resultados da tabela ANOVA e o intervalo de confiança para os parâmetros, possibilitando assim inferir com base no teste acima mencionado.

Tabela 3: Sumarização do modelo ajustado.

Estimativa	Erro Padrão	Estatística t	p-valor
10,9994	0,4897	22,4636	0,000
0,1628	0,0789	2,0639	0,044

Legenda:
[1] Linha 1: Dados referentes a β0
[2] Linha 2: Dados referentes a β1

Tabela 4: Resultados da ANOVA.

GL	SQ	QM	Estatística	p-valor
1	3,6523	3,6523	4,2596	0,044
52	44,5864	0,8574	NA	NA

Legenda:
[1] Linha 1: Dados referentes a β0
[2] Linha 2: Dados referentes a β1
[3] GL: Graus de Liberdade
[4] SQ: Soma de Quadrados
[5] QM: Quadrado Médio
[6] Estatística: F-Snedecor

Tabela 5: Intervalo de Confiança.

α/2 = 2,5%	1-α/2 = 97,5%
10,0168	11,9819
0,0045	0,3211

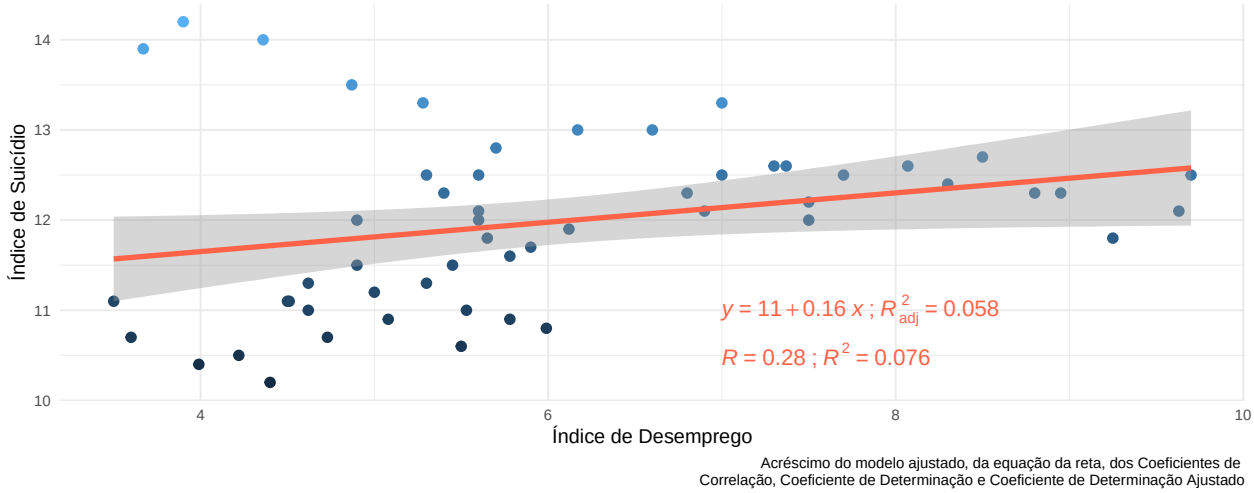
Legenda:
[1] Linha 1: Dados referentes a β0
[2] Linha 2: Dados referentes a β1

Analisando a Tabela 3, que traz os dados sumarizados do modelo ajustado, é possível constatar que tanto $\hat{\beta}_0$ quanto $\hat{\beta}_1$ são significantes para o modelo ajustado, com base no p-valor.

A Tabela 4, que traz os resultados da tabela ANOVA, corrobora com a significância do $\hat{\beta}_1$, pois sendo o p-valor menor que o nível de significância (α) possibilita rejeitar H_0.

A Tabela 5, que traz o Intervalo de Confiança para os parâmetros estimados, mostra que com 95% de confiança é possível afirmar que o verdadeiro valor de $\hat{\beta}_0$ está entre (10,07; 11,98) e que o verdadeiro valor de $\hat{\beta}_1$ está entre (0,004; 0,32).

Figura 5: Reta Ajustada



Acréscimo do modelo ajustado, da equação da reta, dos Coeficientes de Correlação, Coeficiente de Determinação e Coeficiente de Determinação Ajustado

A Figura 5 traz a implementação da reta de regressão ajustada, além da equação da reta e dos Coeficientes de Correlação, Coeficiente de Determinação e Coeficiente de Determinação Ajustado. Tendo em vista o valor do

Coeficiente de Determinação é possível constatar que este não é um bom modelo, pois não explica a variabilidade total dos dados, explicando aproximadamente 7,6% das informações, ou seja, apenas 7,6% das informações referentes ao Índice de Suicídio são explicadas pelo Índice de Desemprego, sendo assim não faz sentido continuar a análise baseado neste modelo.

Conclusão

Embora tenha-se constatado haver uma correlação linear entre as variáveis em análise, tenha-se avaliando os parâmetros do modelo ajustado e os mesmos sendo significativos, com base nas técnicas aprendidas até o momento, se conclui que o modelo estimado não é adequado para explicar a variabilidade total dos dados com base no Coeficiente de Determinação calculado, ou seja, se conclui que o modelo não foi suficiente para explicar significativamente os Índices de Suicídio.