

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

Fernando Bispo, Jeff Caponero

Sumário

Introdução	2
Apresentação	2
Objetivos	2
Análise dos dados	2
Análise Prévia	2
Regressão Linear	4
Estimativas Pontuais	5
Intervalos de Confiança	5
Teste de Significância	5
Conclusão	6

Introdução

O presente relatório tem como objetivo a introdução das técnicas de Regressão Linear Simples e a prática da elaboração de relatórios analíticos fundamentadas na Análise Exploratória de Dados, na resolução de análises conforme os pré-requisitos solicitados para um conjunto de dados.

Apresentação

Será realizada uma análise sobre um levantamento das taxas de desemprego e o índice de suicídios nos EUA para o período de 1950 a 2019, para o qual o índice de suicídios foi calculado para cada 1000 habitantes.

As variáveis contidas no arquivo “desemprego.csv” são:

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

Objetivos

O objetivo dessa análise visa responder aos seguintes tópicos:

- (a) Faça as análises necessárias para verificar se suicídios é função linear do desemprego.
- (b) Obtenha ainda as estimativas das variâncias de β_0 e β_1 .
- (c) Teste a significância do modelo. Que conclusões você chegou com um nível de significância de 5%?
- (d) Obtenha os intervalos de confiança para os parâmetros do modelo com o nível de 95% de confiança. Interprete os resultados.

Análise dos dados

Análise Prévia

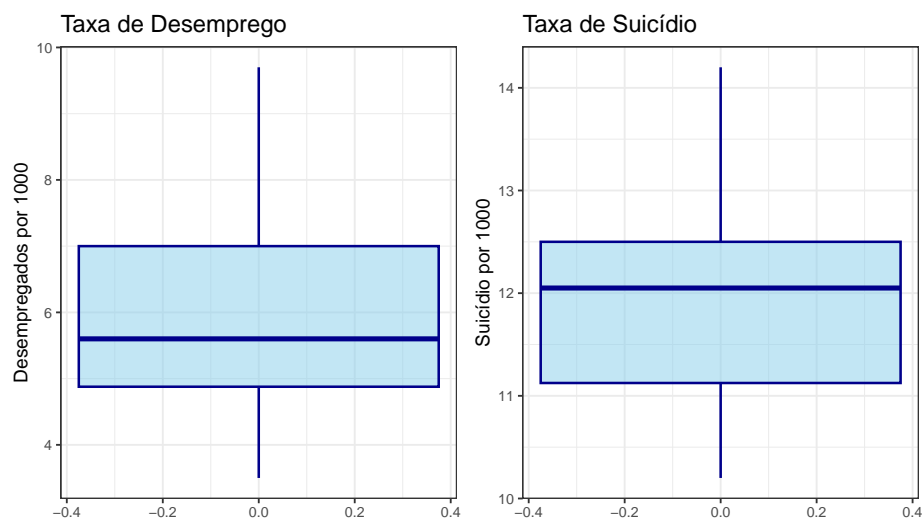
Inicialmente vamos verificar as principais medidas resumo dos dados apresentados e verificar a viabilidade gráfica de realizar uma regressão linear a partir dos dados fornecidos.

Table 1: Medidas resumo das taxas de desemprego e suicídio, nos EUA de 1950 a 2019

	Min	Q1	Med	Média	Q3	Max	D. Padrão	CV
T.Desemprego	3,5	4,87	5,60	6,00	7,0	9,7	1,61	0,27
T.Suicídio	10,2	11,10	12,05	11,98	12,5	14,2	0,95	0,08

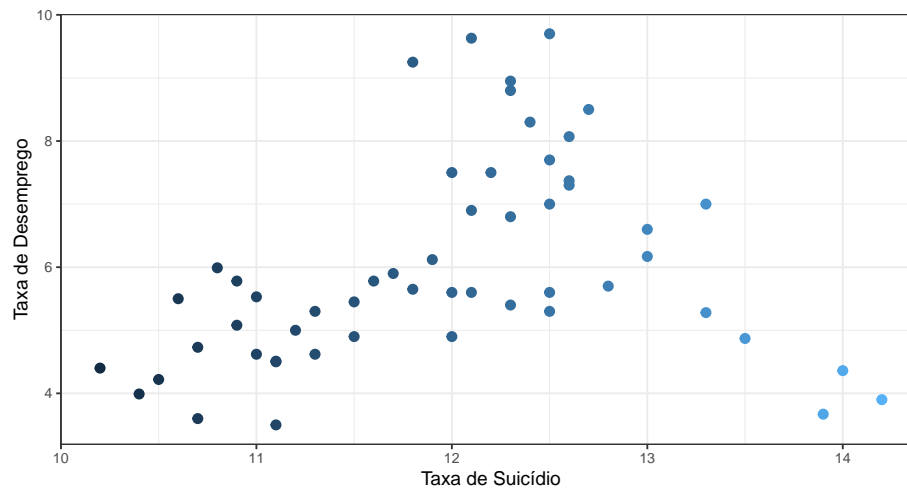
As medidas resumo das taxas de desemprego e suicídio avaliadas não apresetam valores incompatíveis com a análise de regressão linear pretendida.

Figura 1: BoxPlots das taxas de desemprego e suicídio, nos EUA de 1950 a 2019



A análise dos BoxPlots, um pouco mais informativa, uma vez que agora é possível verificar que não há valores notadamente discrepantes que poderiam influir negativamente na regressão linear.

Figura 2: Relação entre Taxa de Desemprego e a Taxa de Suicídio entre 1950 e 2019, nos EUA

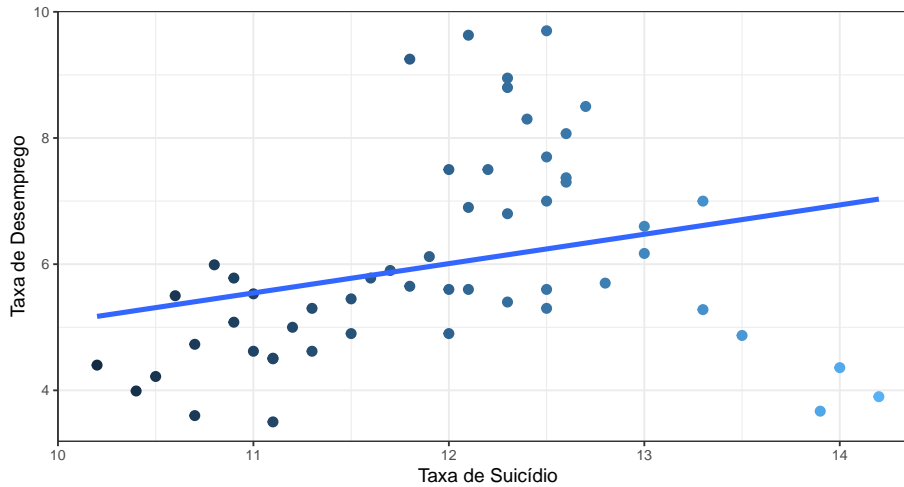


A dispersão de pontos mostrada na Figura 2 não indica visualmente uma tendência linear entre as duas variáveis analisadas, o que pode comprometer todos os resultados seguintes. Admitindo que os dados realmente possam ser explicados por uma regressão linear serão realizados as demais análises.

Regressão Linear

Desta forma, uma possível reta que apresenta a regressão linear dos dados é mostrada na figura a seguir.

Figura 3: Relação entre Taxa de Desemprego e a Taxa de Suicídio entre 1950 e 2019, nos EUA e sua regressão linear.



Estimativas Pontuais

A partir do método de Estimativas de Mínimos Quadrados, pode-se obter uma estimativa pontual para $\hat{\beta}_0 = 0.429$ e para $\hat{\beta}_1 = 0.465$.

Intervalos de Confiança

A partir dos dados é possível obter um intervalo de confiança para as estimativas obtidas. No modelo de regressão linear simples as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ têm distribuição de student, e portanto, pelo método da quantidade pivotal é possível determinar seu intervalo de confiança.

Procedendo com esses cálculos temos que o intervalo de confiança de $\hat{\beta}_0$ é $[-1.356, 2.214]$ e o intervalo de confiança de $\hat{\beta}_1$ é $[0.316, 0.614]$, para um nível de confiança de 95%.

Teste de Significância

Como vimos anteriormente, a avaliação gráfica da dispersão dos dados não indica claramente uma relação linear entre os dados avaliados. Entretanto, pode-se realizar um teste de significância do modelo a fim de melhor compreender quanto adequado é esse modelo.

Para tanto estimou-se o valor do coeficiente de correlação linear dos dados $\hat{\rho} = 0.275$. Em seguida calculou-se o valor da estatística temos que $t = 2.0626$.

Verificou-se que o valor tabelado dessa estatística $t_{(n-1);\alpha/2} = 0.3969$. Por fim, comparando os valores absolutos dessas estatísticas, verificamos que a hipótese de correlação nula deverá ser rejeitada, ao nível de significância de $\alpha = 5\%$.

Conclusão