

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
Transformação dos valores	6
Testes de diagnóstico	7
Conclusão	9

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

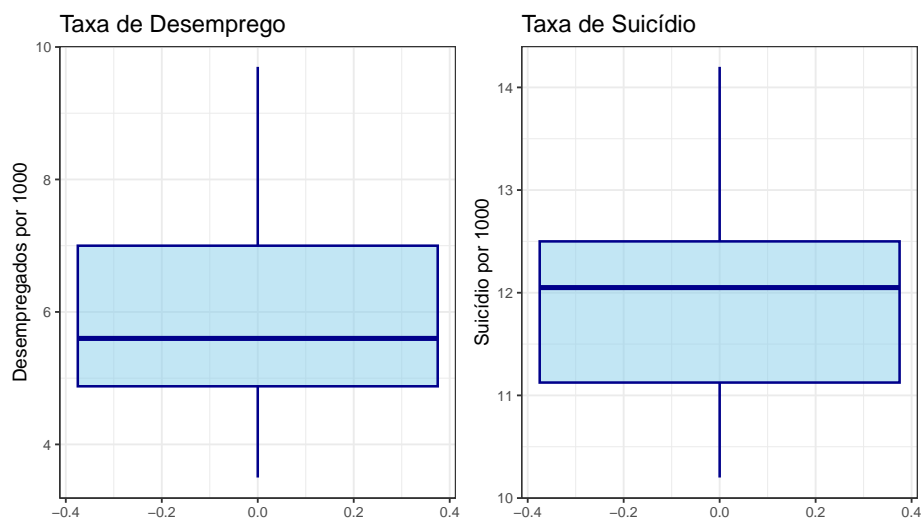
Inicialmente vamos verificar as principais medidas resumo dos dados apresentados e verificar a viabilidade gráfica de realizar uma regressão linear a partir dos dados fornecidos.

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

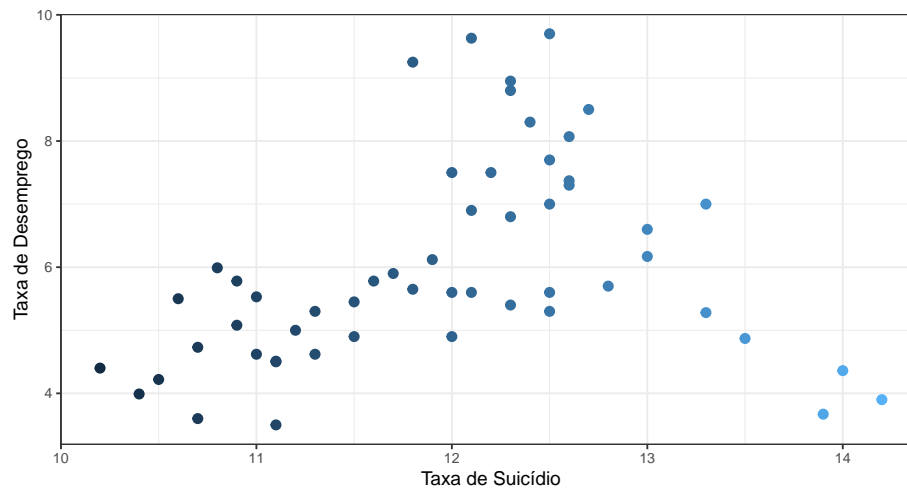
As medidas resumo das taxas de desemprego e suicídio avaliadas não apresetam valores incompatíveis com a análise de regressão linear pretendida, além de mostrarem uma provável normalidade dos dados.

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. A provável normalidade dos dados verificada nas medidas anteriores não parece ter sido afetada pela análise desses gráficos, uma vez que ainda se pode identificar certa simetria nos dados. Entretanto, faz-se necessário a verificação de tal característica, o que será apresentado a seguir.

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

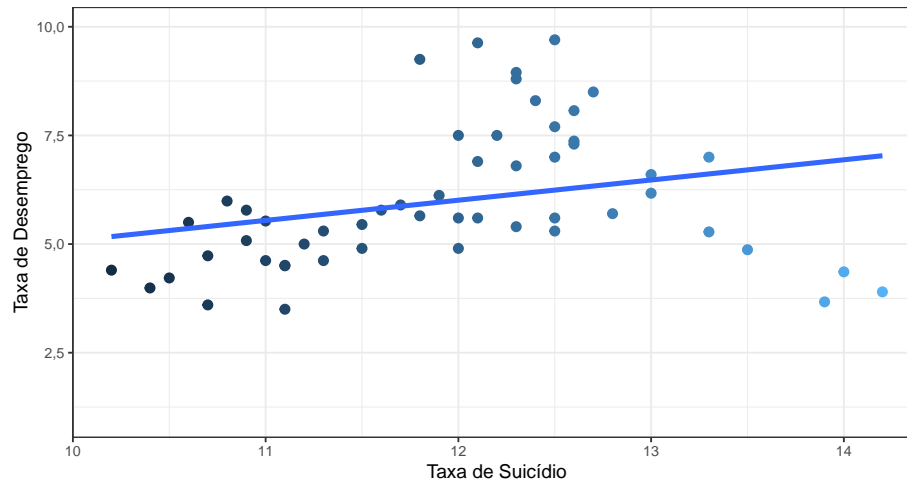


A dispersão de pontos mostrada na Figura 2 não indica visualmente uma tendência linear entre as duas variáveis analisadas, o que pode comprometer todos os resultados seguintes. Admitindo que os dados realmente possam ser explicados por uma regressão linear serão realizados as demais análises.

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 para $\hat{\beta}_0$ 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%. É importante notar que o intervalo calculado de $\hat{\beta}_0$ é bastante largo, compreendendo inclusive valores negativos o que não faz sentido para os dados analisados, logo um intervalo de confiança mais realista para $\hat{\beta}_0$ é $[0, 2.214]$.

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 quão

adequado é esse modelo.

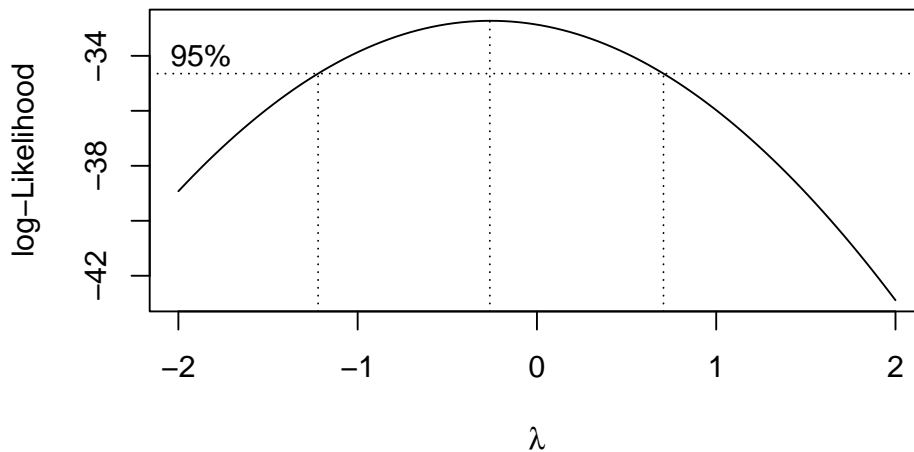
Tomando o método dos mínimos quadrados, estimou-se o valor do coeficiente de correlação linear dos dados $\hat{\rho} = 0.275$. Em seguida calculou-se o valor da estatística temos que $t = 2.0626$. Verificou-se que o valor tabelado dessa estatística $t_{(n-1);\alpha/2} = 0.3969$. Por fim, comparando os valores absolutos dessas estatísticas, verificamos que a hipótese de correlação nula deverá ser rejeitada, ao nível de significância de $\alpha = 5\%$.

Outra forma de avaliar a significância do modelo é realizar uma análise gráfica dos resíduos do modelo.

Transformação dos valores

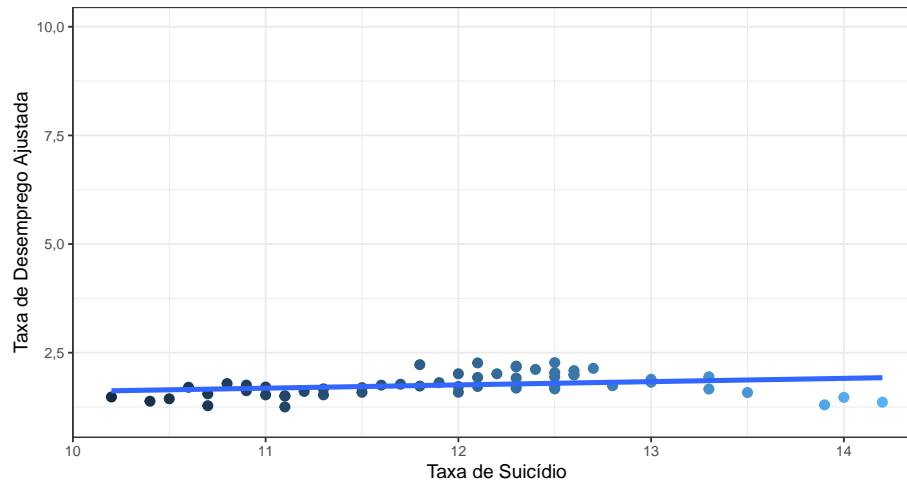
Uma vez que o teste de significância do modelo não apresentou resultado satisfatório, pode-se avaliar a possibilidade de aplicar uma transformação no valores dos dados. Para tanto, pode-se aplicar o método de Box-Cox de avaliação da melhor transformação para os dados.

Figura 4: Escolha de λ na transformação de Box-Cox.



Como se pode verificar pela Figura 4, o valor de λ está próximo ao valor de zero (-0.263). Neste caso, a melhor transformação para os valores de Y é $\log(Y)$, o que levará a estabilização da variância a uma normalização dos dados.

Figura 5: Relação entre Taxa de Desemprego ajustada e a Taxa de Suicídio entre 1950 e 2019, nos EUA e sua regressão linear.



A Figura 5 mostra a dispersão dos dados e o modelo de regressão linear após o ajuste dos valores dos dados. Manteve-se a mesma escala da Figura 3 para evidenciar a transformação, o que leva crer que um melhor ajuste foi obtido.

Desta forma, refazendo as análises anteriores, verifica-se que, pode-se obter novas estimativas para $\hat{\beta}_0^* = 0.856$, com intervalo de confiança de $[0.809, 0.903]$ e $\hat{\beta}_1^* = 0.075$, com intervalo de confiança de $[0.071, 0.079]$, para um nível de confiança de 95%. É importante notar que os novos intervalos calculados são muito mais estreitos e indicam um melhor ajuste do modelo.

O novo coeficiente de correlação linear dos dados $\hat{\rho}^* = 0.273$, com o valor da estatística $t = 2.0464$ e $t_{(n-1);\alpha/2} = 0.3969$, logo, a nova hipótese de correlação nula deverá ser rejeitada, ao nível de significância de $\alpha = 5\%$.

Testes de diagnóstico

Pode-se ainda utilizar um conjunto de testes de diagnóstico para confirmar este novo teste de significância. Como:

- Teste de Kolmogorov-Smirnov
- Teste de Shapiro-Wilks
- Teste de Goldfeld-Quandt
- Teste de Breush-Pagan

- Teste de Park
- Teste F para linearidade
- Teste para avaliação da independência dos resíduos

Teste de Kolmogorov-Smirnov

Avalia o grau de concordância entre a distribuição de um conjunto de valores observados e determinada distribuição teórica. Consiste em comparar a distribuição de frequência acumulada da distribuição teórica com aquela observada. Realizado o teste obteve-se um p-valor de 0.88, o que inviabiliza rejeitar a hipótese de que haja normalidade entre os dados, com um grau de confiabilidade minimamente razoável.

Teste de Shapiro-Wilks

O teste de Shapiro-Wilks é um procedimento alternativo ao teste de Kolmogorov-Smirnov para avaliar normalidade. Realizado o teste obteve-se um p-valor de 0.627, o que, semelhantemente, inviabiliza rejeitar a hipótese de que haja normalidade entre os dados, com um grau de confiabilidade minimamente razoável.

Teste de Goldfeld-Quandt

Esse teste envolve o ajuste de dois modelos de regressão, separando-se as observações das duas extremidades da distribuição da variável dependente. Realizado o teste obteve-se um p-valor de 0.054, o que inviabiliza rejeitar a hipótese de que haja homocedasticidade entre os dados, com um grau de confiabilidade de 95%. Entretanto, como o p-valor obtido é próximo do necessário para a rejeição da hipótese nula, cabe um novo teste para a confirmação do resultado obtido.

Teste de Breush-Pagan

Esse teste é baseado no ajuste de um modelo de regressão em que a variável dependente é definida pelos resíduos do modelo de interesse. Se grande parte da variabilidade dos resíduos é explicada pelo modelo, então rejeita-se a hipótese de homocedasticidade. Realizado o teste obteve-se um p-valor de 0.004, desta forma deve-se rejeitar a explicação dada pelo teste, assim, não se pode rejeitar a hipótese de que haja homocedasticidade entre os dados, com um grau de confiabilidade minimamente razoável.

Teste de Park

Esse teste é baseado no ajuste de um modelo de regressão em que a variável dependente é definida pelos quadrados dos resíduos do modelo de interesse. Nesse caso, se β_1 diferir significativamente de zero, rejeita-se a hipótese de homocedasticidade. O valor de β_1 obtido no teste foi de 0.037 com p-valor de 0.002. Por esse teste não se deve rejeitar a hipótese de homocedasticidade, com confiabilidade de 95%.

Teste F para linearidade

O teste da falta de ajuste permite testar formalmente a adequação do ajuste do modelo de regressão. Neste ponto assume-se que os pressupostos de normalidade, variância constante e independência são satisfeitos, como demonstrado pelos testes realizados. A ideia central para testar a linearidade é decompor SQ_{Res} em duas partes: erro puro e falta de ajuste que vão contribuir para a definição da estatística de teste F. Realizado o teste obteve-se um valor de p-valor igual a 0.042, o que demanda a rejeição da hipótese que há uma relação linear entre as variáveis.

Teste para avaliação da independência dos resíduos

Tendo em vista, o resultado obtido no teste anterior esse teste pode esclarecer ainda mais o ajuste do modelo.

O teste para avaliação da independência dos resíduos é utilizado para detectar a presença de autocorrelação provenientes de análise de regressão. Realizando o teste obteve-se um valor de p-valor aproximadamente igual a 0, indicando que se deve rejeitar a hipótese que não existe correlação serial entre os dados, com uma confiança de 95%.

Conclusão

Em uma análise preliminar verificou-se que os dados apresentavam certa normalidade, mas graficamente não podia se afirmar que estabelecessem uma relação linear. Aplicando-se o método dos quadrados ao modelo de regressão linear proposto, verificou-se que o mesmo não obteve o resultado desejado em um teste de significância. Por conta deste teste, propôs-se realizar uma transformação dos dados conforme avaliação proposta por Box-Cox. A transformação logarítmica dos dados levou a uma normalização e homocedasticidade conforme verificada pelos testes realizados. Entretanto, a transformação que já havia obtido um resultado insatisfatório no novo teste de significância, não foi capaz de obter um melhor resultado na avaliação de sua linearidade por um teste F. Assim, tentou-se avaliar se haveria uma dependência linear entre os resíduos do modelo, o que não se provou aceitável.

Desta forma, embora a transformação tenha garantido as premissas para aplicação de uma regressão linear essa não foi suficiente para explicar o comportamento dos dados e isto não se deveu a uma possível correlação serial dos resíduos.