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 a 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.

Apresentação

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

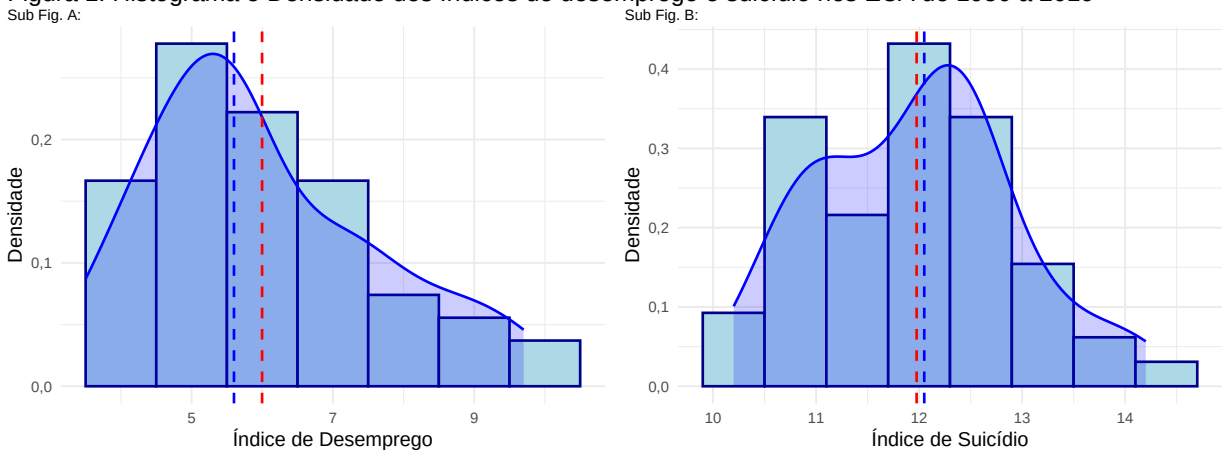
	DESEMP	SUIC
Mín	3,50	10,20
Q1	4,87	11,10
Med	5,60	12,05
Média	6,00	11,98
Q3	7,00	12,50
Máx	9,70	14,20
Desv.padrão	1,61	0,95
CV	0,27	0,08
Assimetria	0,63	0,22
Curtose	-0,50	-0,56

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

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.

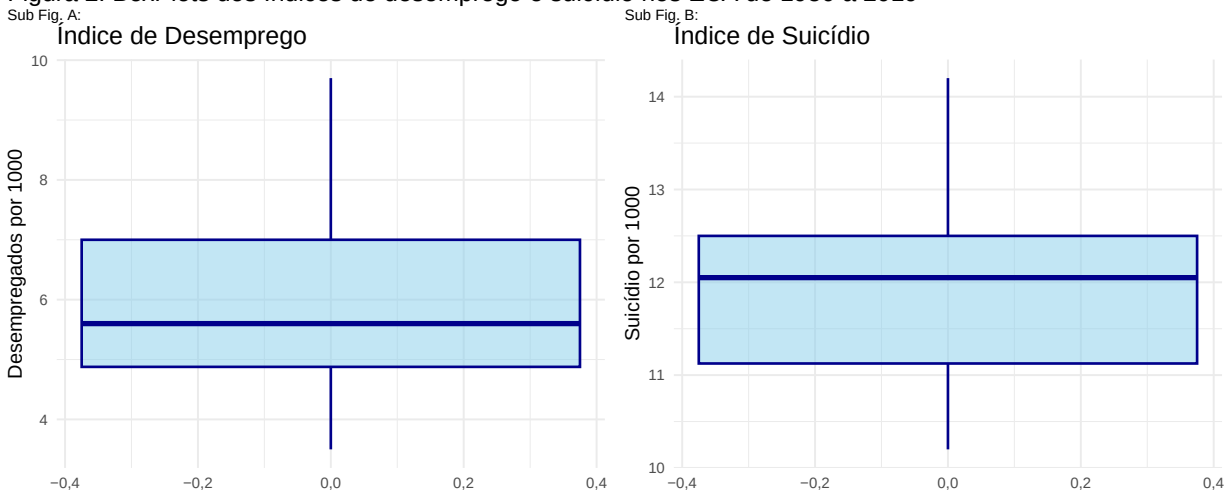
Figura 1: Histograma e Densidade dos Índices de desemprego e suicídio nos EUA de 1950 a 2019



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

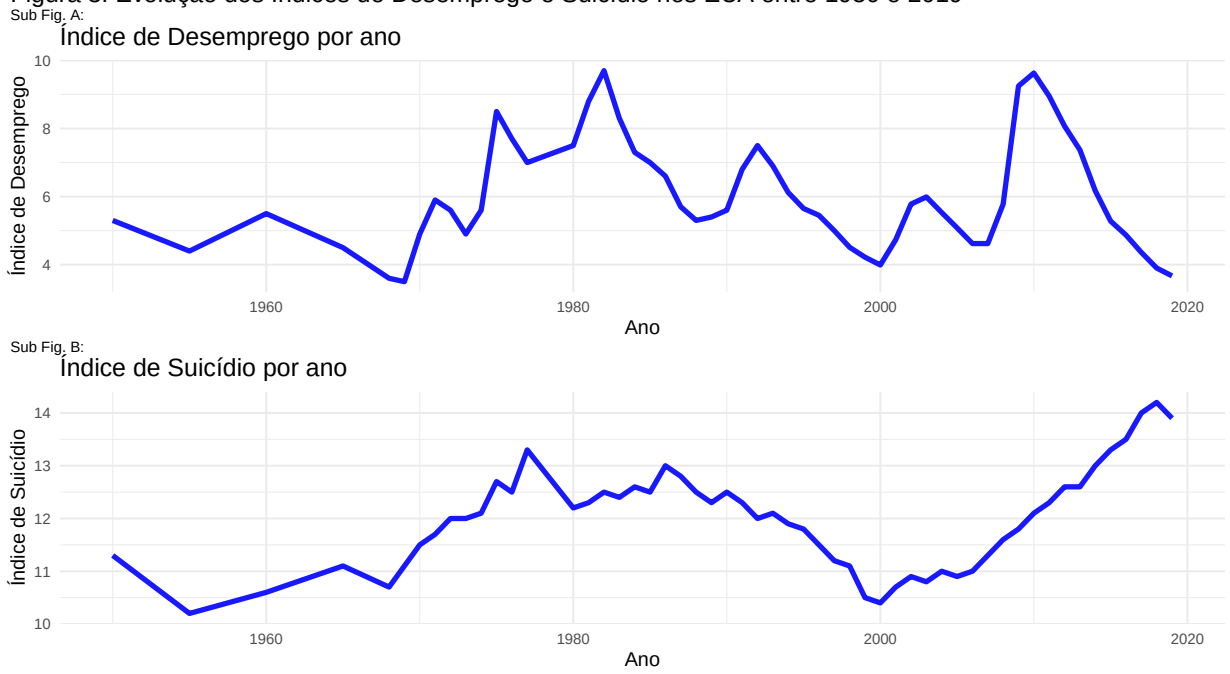
Conforme identificado na Tabela 1, a Figura 1 traz a representação gráfica das conclusões realizadas para os dados, no que diz respeito a assimetria e curtose.

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 que não há valores notadamente discrepantes (*outliers*) que poderiam influir negativamente na regressão linear.

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

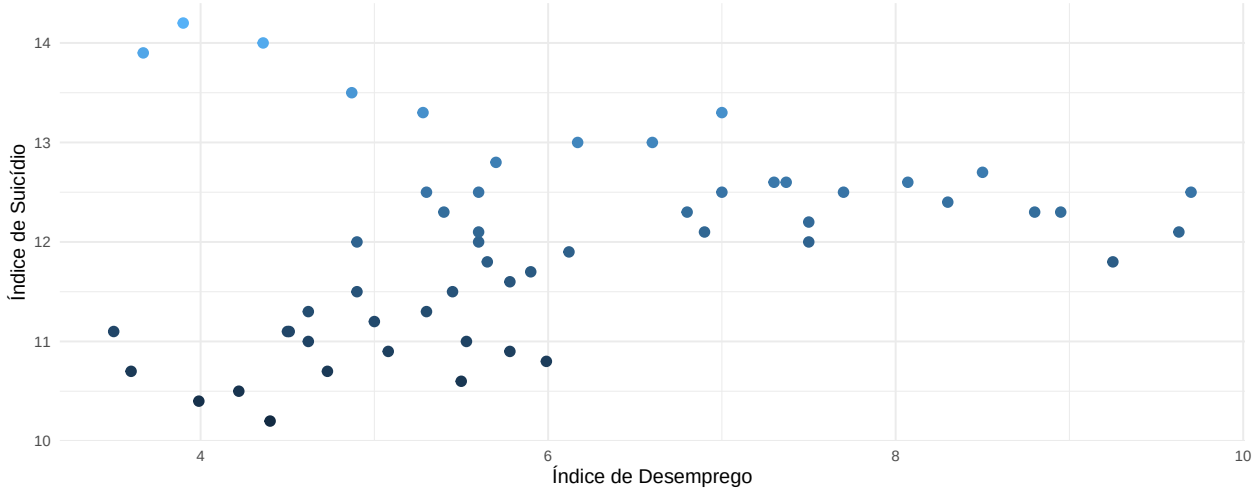


Através da Figura 3 é possível avaliar a evolução dos índices em analise ao longo do tempo, em que é possível identificar uma queda no Índice de Desemprego nos anos finais em que os dados foram coletados, contudo o Índice de Suicídio está apresentando um sinal de queda após um longo período de alta.

A fim de analisar a relação entre as variáveis 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.

A fim de se ter uma melhor percepção acerca dessa relação, se construiu a Figura 4 para avaliação dessa relação entre as variáveis.

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

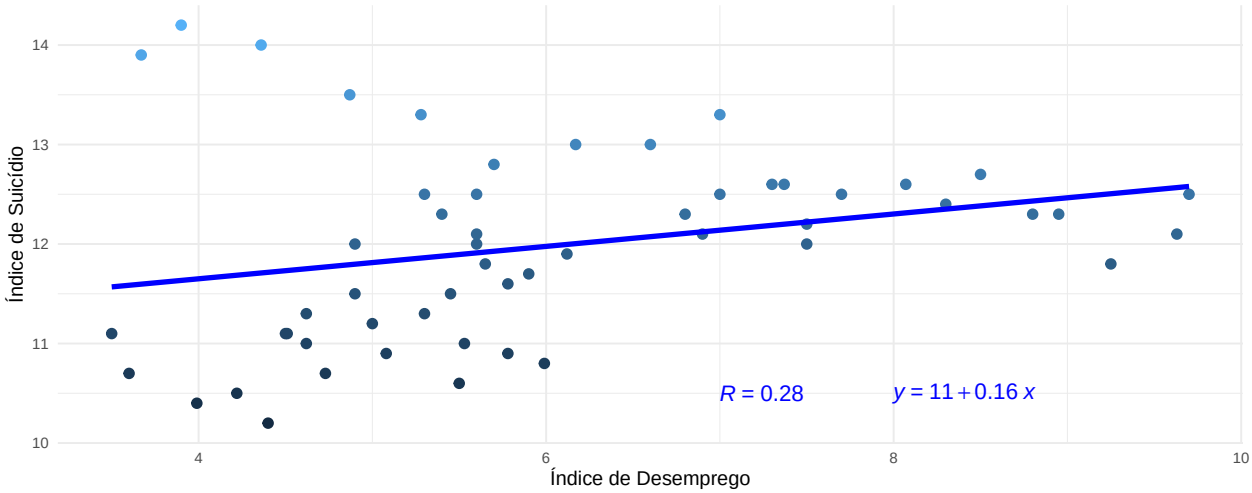


Corroborando com o Coeficiente de Correlação ($\hat{\rho}$), não é possível notar uma aparente correlação entre as variáveis, o que pode comprometer todos os resultados seguintes.

Regressão Linear

Ainda assim, admitindo que os dados dos Índices de Suicídio realmente possam ser explicados por uma regressão linear foi construída a Figura 5, em que se acrescentou a reta de Regressão aos dados em análise.

Figura 5: Relação entre Índice de Desemprego e a Índice de Suicídio entre 1950 e 2019, nos EUA



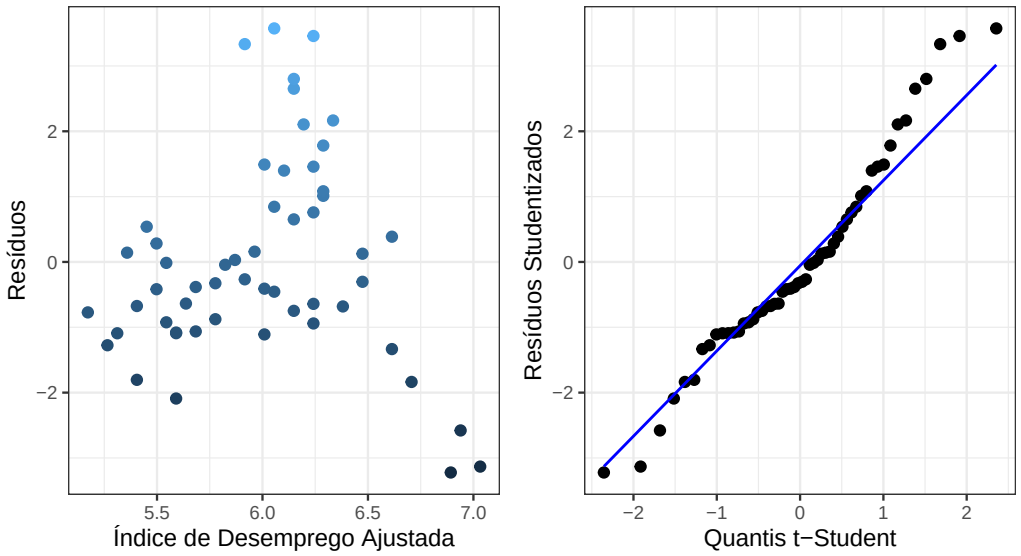
Análise dos Resíduos

A partir do modelo proposto foram feitas as análises gráficas apresentadas na Figura 6.

Figura 6: Análise de resíduos do modelo de regressão proposto.

Sub Fig. A:

Sub Fig. B:



A Figura 6 permite observar a falta de homocedasticidade tanto pelo gráfico dos valores ajustados pelo modelo e os resíduos, quanto pelo gráfico dos quantis t-Student. No primeiro caso, não se observa que os pontos fiquem em torno do zero e no segundo, não se ajustam a reta.

Estimativas Pontuais

A partir do método de Estimativas de Mínimos Quadrados, pode se obter uma estimativa pontuais 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 analizados, logo um itervalo de confiança mais realista para $\hat{\beta}_0$ é [0, 2.214].

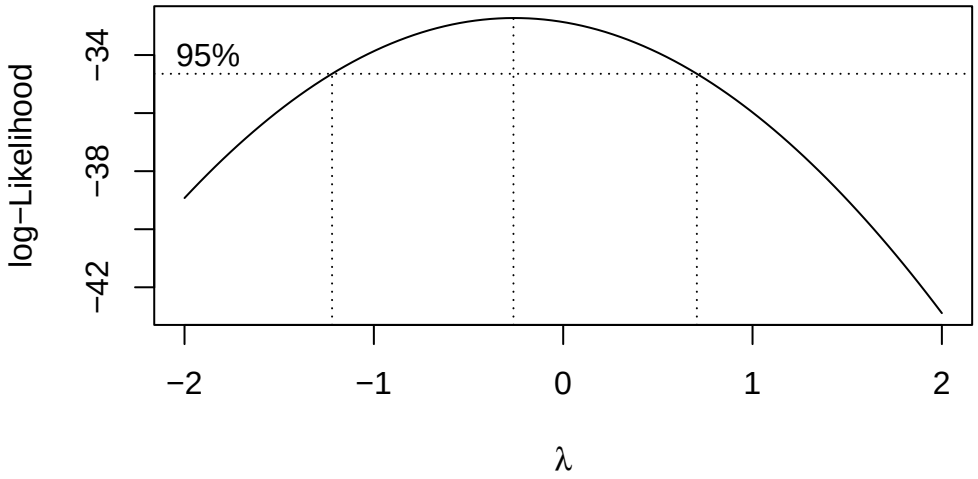
Teste de Significância

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 siguinificância do modelo é realizar uma análise gráfica dos resíduos do modelo.

Transformação dos valores

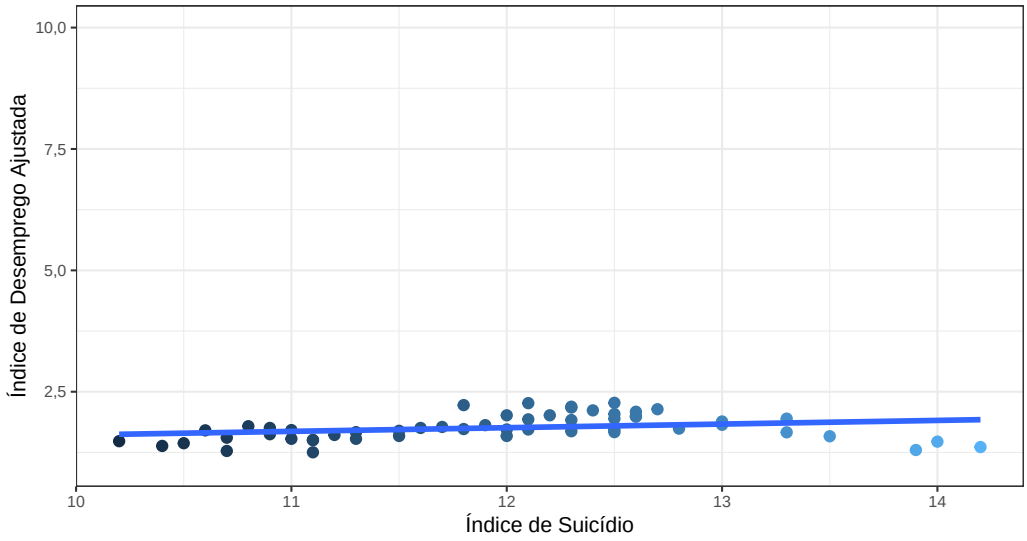
Uma vez que o teste de significância do modelo não apresentou resultado satisfatório, pode-se avaliar a possibilidade de aplicar uma transformação no valores dos dados. Para tanto, pode-se aplicar o método de Box-Cox de avaliação da melhor transformação para os dados.

Figura 5: Escolha de λ na transformação de Box-Cox.



Como se pode verificar pela Figura 5, 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 6: Relação entre Índice de Desemprego ajustada e a Índice de Suicídio entre 1950 e 2019, nos EUA e sua regressão linear.



A Figura 6 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 análise 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 demanda 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 hipotese 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 não é explicada pelo modelo, então rejeita-se a hipótese de homocedasticidade. Realizado o teste obteve-se um p-valor de 0.004, desta foram deve-se rejeitar a hipótese de que haja homocedasticidade entre os dados, com um grau de confiabilidade de 95%.

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 demosntrado pelos testes realizados. A ideia central para testar a linearidade é decompor SQRes 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 valore 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 aproximadadente igual a 0, indicando que se deve rejeitar a hipotese 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 realção 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 logarítimica dos dados levou a uma normalização e homocedasticidade conforme verificada pelos testes realizados. Entretanto, a trasformação 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.