Revista Gestão Industrial
ISSN: 1808-0448
http://periodicos.utfpr.edu.br/revistagi

Correlação e Regressão Linear de Variáveis que interferem no Produto Interno Bruto do Brasil: Uma Análise Estatística de Dados

Guilherme Mateus Kremer
gkremer@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná (UTFPR), Ponta Grossa, Paraná, Brasil

Carolina Deina
caroldeina@gmail.com
Universidade Tecnológica Federal do Paraná (UTFPR), Pato Branco, Paraná, Brasil

Hugo Siqueira
hugosiqueira@utfpr.edu.br
Universidade Tecnológica Federal do Paraná (UTFPR), Ponta Grossa, Paraná, Brasil

RESUMO

Este trabalho tem como objetivo analisar a correlação de 3 variáveis quantitativas e, a partir delas, elaborar uma equação de regressão linear para realizar previsão do Produto Interno Bruto (PIB) do Brasil, utilizando as variáveis independentes Expectativa de Vida e População do Brasil. Para isso, coletou-se séries temporais de dados entre o período de 1960 até 2016, e com o auxílio do *Software IBM SPSS Statictic* e buscou-se analisar a interferência dos dados das variáveis independentes sobre a dependente, através de ferramentas de análises estatísticas como testes de normalidade de dados, análises descritivas de variáveis, correlação de variáveis e regressão linear. Como resultados, foi verificado a possibilidade de utilizar o método de regressão linear obtendo uma equação de reta para a previsão do PIB brasileiro para os anos de 2017 e 2020, utilizando as variáveis explicativas propostas. Essa possibilidade foi constatada, pois a população e expectativa de vida apresentaram correlações positivas e significativas em relação ao PIB, sendo que o percentual de acerto foi de 76,20%, valor esse, considerado aceitável para as bases de dados selecionadas. Dessa forma, concluiu-se que o PIB cresce ou descresse à medida em que ocorre mudanças nas variáveis citadas.

PALAVRAS-CHAVE: Análise de dados estatísticos. Correlação. Regressão linear. Produto interno bruto.

INTRODUÇÃO

O Produto Interno Bruto (PIB) é composto pela produção total de bens e serviços de um país, sendo soma dos valores gerados pela agricultura, indústria e serviços (IBGE, 2018). No cálculo da produção total são descontados os gastos com insumos utilizados durante o próprio processo produtivo ativo. É o indicador econômico geral da economia de um país e de toda a sua produção, sendo um dos mais utilizados mundialmente para quantificar a atividade econômica da região. A mensuração do PIB de um país é realizada pelo órgão competente a cada, seguindo a metodologia da Organização das Nações Unidas (ONU). No caso do Brasil, o responsável é o Instituto Brasileiro de Geografia e Estatísticas (IBGE) (GOVBR, 2016).

O cenário econômico, político e social influencia consideravelmente no resultado de um PIB (IBGE, 2018). Durante a década de 1990, o PIB brasileiro sofreu grandes oscilações por conta do instável cenário de hiperinflação e instabilidade, os quais foram fortemente mitigados com a implementação do Plano Real em 1994, através principalmente pela mudança do regime cambial (OLIVEIRA; TUROLLA, 2003). Quando a inflação entrou em queda, o poder de compra da população foi afetado ao mesmo tempo que o acréscimo salarial estimulou o consumo, levando ao aquecimento da economia nacional (BASTOS *et al.,* 2015).

No período de 1997 a 2000, o país foi marcado por crises externas e internas que afetaram o desempenho econômico nacional. A Crise Asiática, em 1997, fez com que parte dos países asiáticos perdessem força econômica temporária por conta da desvalorização cambial e perda de reservas. Isto atingiu negativamente as bolsas de valores de todo globo. A Crise Brasileira em 1999, ano em que o câmbio deixou de ser fixo e passou a ser flutuante devido a acentuada queda das reservas monetárias, acarretou uma taxa de crescimento do PIB de apenas 0,3% (BASTOS *et al.,* 2015). Esses exemplos ao longo de um período curto da história brasileira demonstram como os diversos fatores podem interferir no desempenho no PIB de um país (RIBEIRO *et al.,* 2010).

Além disso, na década de 1950, a expectativa de vida ao nascer no Brasil era de menos de 50 anos e passou para 74,8 anos em 2013, sendo esta uma média geral do país o que não condiz com os números reais de todas as regiões. Essas taxas cresceram ao mesmo tempo em que as taxas de mortalidade infantil diminuíram (IBGE, 2018). Além disso, no país há uma escassez de pesquisas em relação ao estudo de envelhecimento populacional e expectativa de vida.

A expectativa está ligada diretamente ao envelhecimento da população brasileira e da melhoria nas condições de vida, como medidas de saúde pública. Em estudos anteriores que tiveram com base os anos de 1998 a 2008, pode-se constatar que o nível de saúde da população classificada entre ruim, regular e boa, não sofreu consideráveis mudanças entre pessoas do sexo masculino. Entretanto, no sexo feminino houve uma mudança considerável, aumentando na classificação boa, mantendo-se na regular e diminuindo na ruim. Isso indica uma influência na expectativa de vida e interfere na população e na taxa de crescimento do país.

O Brasil possui um grande contingente populacional, ultrapassando os 200 milhões de habitantes em 2016. A população cresceu vertiginosamente desde o primeiro censo em 1872, quando havia 10 milhões de habitantes. O aclive de

crescimento ocorreu após os anos de 1950, quando houve as maiores taxas da série histórica (TIBULO *et al.,* 2012; CAMARGOS; GONZAGA, 2015).

Considerando que possa existir uma relação entre expectativa de vida e população com o PIB brasileiro, este trabalho tem por objetivo investigar a relação entre essas variáveis e propor uma equação de reta através do método de regressão linear, o qual possibilitará obter uma previsão sobre o PIB brasileiro para os anos de 2017 e 2020.

Ressaltamos que outros estudos investigaram a relação entre a variável dependente com as independentes como (Rattanametawee; Leenawong; Netisopakul, 2016; Mandal; Batina; Chen, 2018; Dião *et al*., 2018) que buscaram, respectivamente, desvendar os efeitos de eventos especiais para vendas de carros; a influência de gênero, educação e saúde no crescimento econômico; influência de fatores socioeconômicos e políticos nas emissões de Óxido de Nitrogênio. Nestes foram constatadas evidências de relação entre as variáveis, sendo que qualquer mudança que possa ocorrer nas variáveis causais investigadas afetaram a variável dependente.

Portanto, há evidências de que tal relação que possa existir entre as séries temporais do PIB, expectativa de vida e população, as quais alteram cenários como o de projeção dessas variáveis.

Para a investigação da relação entre as 3 séries temporais propostas nesse estudo, é realizada uma análise de estatística descritiva dos dados, verificando se a sua distribuição histórica apresenta características de normalidade. Após, serão aplicadas as metodologias adequadas para analisar a relação dessas variáveis e qual a interferência da Expectativa de Vida e População sob o Produto Interno Bruto, para então propor uma equação através do método de regressão linear e realizar as projeções para o ano de 2017 até 2020.

Nas próximas seções será apresentado o referencial teórico a respeito dos métodos estatísticos utilizados, a metodologia para coleta e tratamento dos dados, bem como os resultados e conclusões encontrados.

REFERENCIAL TEÓRICO

Esta Seção tem como objetivo embasar teoricamente as ferramentas utilizadas para uma análise de dados estatísticos. O foco desse referencial está nas características dos dados da série histórica utilizada, ou seja, das 3 variáveis quantitativas. Para oferecer esse embasamento, são apresentadas referências quanto à análise descritiva, testes de normalidade, correlação e seus critérios e regressão linear e seus critérios.

Segundo Ribas e Vieira (2011), para a realização dos testes é indispensável que uma exploração inicial dos dados, para se analisar o cumprimento dos pressupostos de cada ferramenta. Caso contrário, pode haver um comprometimento dos resultados.

Para realização do teste de correlação, deve-se primeiramente analisar os dados para estabelecer o tipo de teste a ser aplicado. Métodos paramétricos exigem o teste de correlação de Pearson e os não-paramétricos, de correlação de Spearman, por exemplo. A aplicação da regressão linear múltipla deve obedecer a

pressupostos de linearidade, nenhum ou poucos *outliers* e, também, não presença de multicolinearidade.

Segundo Triola (1999), pela natureza de seus dados, as variáveis podem ser divididas em quantitativas, que consistem em um conjunto de dados que representam contagens ou medidas, e qualitativas ou categóricas, que consistem em um conjunto de dados subjetivos que podem ser separados em diferentes categorias que se distinguem por alguma característica não-numérica.

A variável dependente é conhecida por ser o objetivo do que se pretende analisar, a qual pode sofrer influências de outras variáveis explicativas (causais), conhecidas por variáveis independentes (HAIR et al., 2009).

Segundo Reis e Reis (2002), a análise descritiva é a fase inicial de um estudo de dados coletados. Os métodos dessa avaliação têm como objetivo organizar, resumir e descrever os aspectos importantes das características observadas nos dados ou comparar tais características entre dois ou mais conjuntos. Segundo os mesmos autores, a descrição dos dados também tem como objetivo identificar anomalias e dispersões que não seguem a tendência geral do restante do conjunto (*outliers*).

As ferramentas descritivas são os muitos tipos de gráficos e tabelas e medidas de síntese como porcentagens, índices e médias. Ao aplicar esses métodos aos dados, perde-se informação por conta da condensação, mas, ao mesmo tempo, há um ganho muito maior com relação a clareza da interpretação (REIS; REIS, 2002).

TESTE DE NORMALIDADE DOS DADOS

É de grande relevância realizar uma inspeção inicial nos dados para verificar se as variáveis exibem distribuição adequada para a realização de testes paramétricos ou não paramétricos. Quando se utiliza dados não paramétrico em teste paramétrico, os resultados não exibem precisão (RIBAS; VIEIRA, 2011).

Existem diversas formas de análise de normalidade de dados, entre eles, os testes de Kolmogorov-Smirnov e Shapiro-Wilk. Para ambas as formas expostas neste estudo, segundo Ribas e Vieira (2011), as hipóteses empregadas são:

H_0: Os dados exibem distribuição normal.

H_1: Os dados não exibem distribuição normal.

A rejeição de H_0, ao encontrar uma significância menor que 0,05, sugere que os dados não são provenientes de uma distribuição normal. Se o teste não rejeitar a normalidade, ou seja, se a significância obtida for maior que 0,05, há evidência suficiente para o emprego seguro de um procedimento paramétrico que suponha normalidade (CERVEIRA; SELLITTO, 2015; VIANA *et al.,* 2018).

Kolmogorov-Smirnov e Shapiro-Wilk

Na avaliação de amostras pequenas, geralmente com menos de 30 casos, usa-se o teste de Kolmogorov-Smirnov para a determinação da normalidade ou da não normalidade dos dados. O procedimento para a avaliação da normalidade dos dados segue o exposto acima (MIOT, 2017). Na avaliação de amostras grandes,

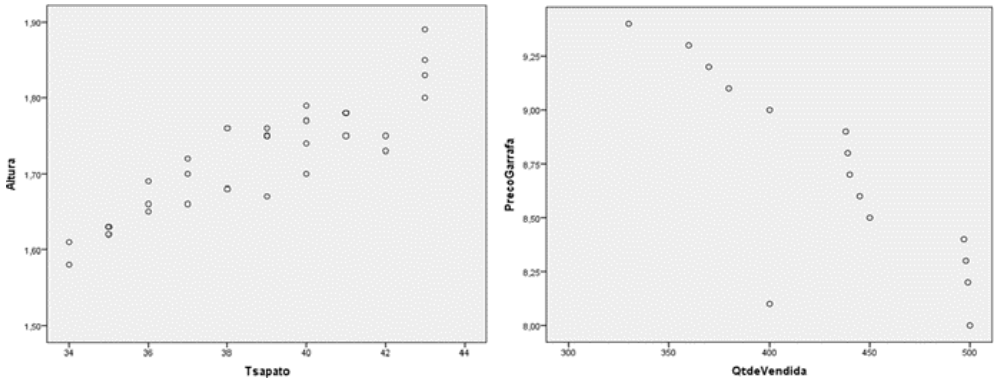
geralmente com mais de 30 casos, usa-se o teste de Shapiro-Wilk para condições semelhantes (RAZALI; WAH, 2011).

CORRELAÇÃO DE VARIÁVEIS

Duas variáveis possuem correlação quando apresentam algum relacionamento entre elas (RICARDO; MEDEIROS; SALAS, 2018). Isto pode ser observado usando um diagrama de dispersão, um gráfico de dados emparelhados (x, y) com eixo x horizontal e o eixo y vertical. Se os pontos do diagrama apresentam um certo padrão, pode-se concluir que há relação entre as variáveis (TRIOLA, 1999).

Ademais, quando duas variáveis apresentam certa correlação, esta pode ser tanto negativa quanto positiva. A correlação negativa acontece quando as variáveis são inversamente proporcionais, ou seja, quando o valor de uma aumenta, o da outra diminui. A positiva acontece quando as variáveis são diretamente proporcionais (FERREIRA, 2013). Os conceitos de correlação positiva e negativa podem ser observados no primeiro e segundo gráfico, respectivamente, na Figura 1.

Figura 1 - Gráficos de correlação



Fonte: Adaptado IBM SPSS Statistic

Para realização de um teste de correlação, primeiramente deve-se definir se os dados da amostra a ser utilizada são de natureza paramétrica ou não-paramétrica. A partir dessa determinação, se encaminha o restante do estudo: teste de Pearson, caso os dados apresentem normalidade; teste de Spearman, na condição de que os dados não possuam normalidade (PILATTI; PICININ; NASCIMENTO, 2017; MIOT, 2017).

Correlação de Pearson e Spearman

O teste de correlação de Pearson é utilizado quando se pretende avaliar se existe correlação entre duas variáveis com dados distribuídos parametricamente, e, a força dessa relação (MIOT, 2017).

Segundo Figueiredo e Silva (2009), a correlação por esse método acarreta o surgimento do coeficiente de correlação de Pearson®, uma medida de associação linear entre variáveis. Esse coeficiente varia entre valores de -1 e 1, e seu resultado indica se a relação entre as variáveis é positiva ou negativa, e, ainda a força dessa

relação. Para Dancey e Reidy (2005) a classificação para esse valor se dá por: r = 0,10 até 0,30 (fraco); r = 0,40 até 0,6 (moderado); r = 0,70 até 1 (forte).

O teste de correlação de Spearman segue os mesmos parâmetros de avaliação que o de Pearson, porém esse tipo de correlação é aplicado para uma série de dados com distribuição não-normal, ou seja, não-paramétrico (PILATTI; PICININ; NASCIMENTO, 2017).

REGRESSÃO

Torres Júnior, Nascimento e Souza (2006) define que o objetivo de uma considerável parcela dos cálculos de regressão é investigar se as variáveis estão relacionadas deterministicamente. Afirmar que x e y estão relacionados dessa forma em um estudo, significa dizer que o conhecimento do valor de x implica no conhecimento do valor de y.

Para a aplicação de uma regressão, seja simples ou múltipla, primeiramente deve-se testar os pressupostos necessários, como pouca ou nenhuma presença de *outliers*, linearidade e presença de não multicolinearidade, e depois, realizar o teste de regressão propriamente dito (DEVORE, 2006).

Outliers

Causados por registro equivocado dos dados ou presentes no próprio fenômeno analisado, os *outliers* são identificados como valores excessivamente reduzidos ou elevados quando comparados com o restante dos elementos. Estudos de amostras contendo esses pontos sobressalentes, podem apresentar resultados distorcidos substancialmente (RIBAS; VIEIRA, 2011).

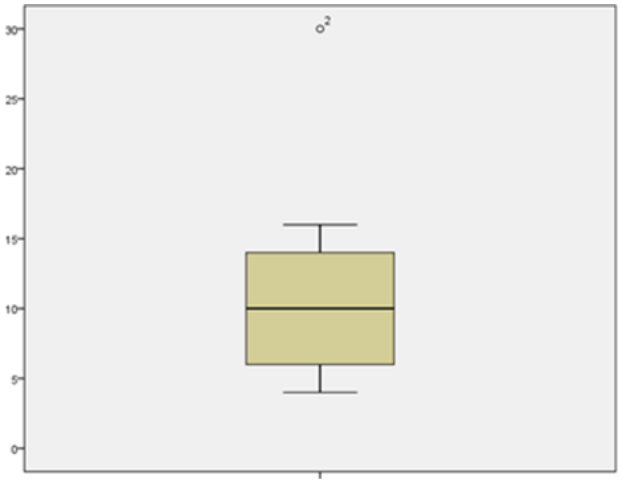
Ainda segundo Ribas e Vieira (2011), a chance de se incluírem casos extremos no estudo é ampliada quando o tamanho da amostra aumenta, mas, nesse caso restrito, não é necessária sua remoção, pois constituem de observações legítimas da população. Outros casos também podem gerar *outliers* que não interfiram no resultado esperado de modo que a retirada ou não desses pontos deve ser analisada separadamente em cada situação.

Existem inúmeros meios de identificar a presença de *outliers* em uma amostra pesquisada. Uma das técnicas utilizadas é a montagem do boxplot (diagrama de caixa) referente aos dados e análise de pontos que ocasionalmente possam se alocar para fora do mesmo, para que se possa concluir em cada caso se devem ou não ser removidos da amostra original (SILVA JÚNIOR; OLIVEIRA, 2005). A Figura 2 exemplifica um diagrama de caixa com a presença de um *outlier*, usando um exemplo aleatório de dados.

Linearidade

Ribas e Vieira (2011), mencionam que a linearidade exige que os dados sejam aglomerados seguindo uma linha reta. Isso não especifica que em determinado estudo os dados necessitam ser perfeitamente lineares, mas sim, que não devem exibir sinais claros de não linearidade. Conforme apresentado na Figura 3, os dados

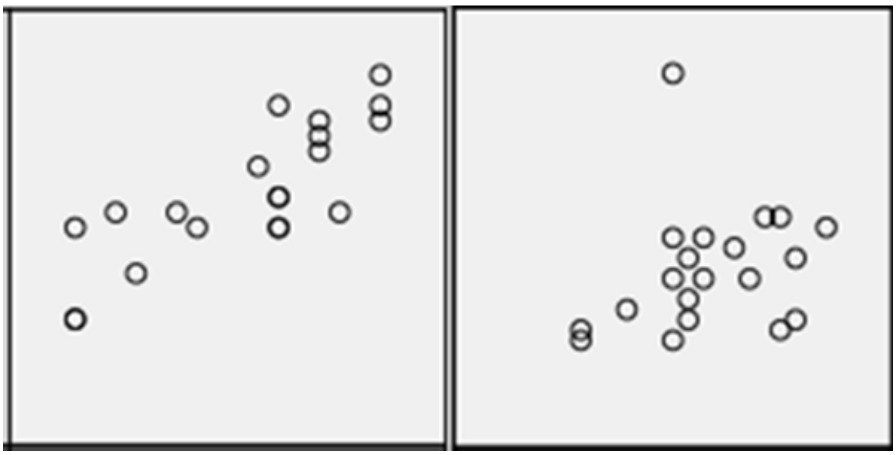
Figura 2 - Boxplot



Fonte: Adaptado IBM SPSS Statistic

do primeiro gráfico (a) de dispersão apresentam linearidade, e os do segundo gráfico, apresentam uma certa não-linearidade, pois não seguem um padrão linear.

Figura 3 - Gráficos de dispersão (a) com linearidade e (b) sem linearidade



(a) (b)

Fonte: Adaptado IBM SPSS Statistic

Multicolinearidade

A multicolinearidade é a condição existente quando duas ou mais variáveis independentes são fortemente correlacionadas (BURGEL; ANZANELLO, 2018). Ribas e Vieira (2011) afirmam que a presença desse parâmetro pode distorcer a interpretação dos resultados, pois, se duas variáveis foram altamente correlacionadas, elas podem estar mensurando a mesma característica, não sendo possível identificar qual a relevância de cada uma na interpretação do estudo. Se o valor da correlação existente for superior a 0,9, considera-se a presença de multicolinearidade.

Regressão Linear Múltipla

Deve-se aplicar a regressão linear simples, quando deseja-se saber o quanto uma certa variável independente influencia em uma variável dependente, ou seja, quando pretende-se determinar a relação existente entre elas (PASSOS *et al.,* 2012).

Para uma dada regressão linear simples, a partir de um estudo de uma coleção de dados amostrais emparelhados, a equação de regressão $y = a + bx$ descreve a relação entre as duas variáveis, *x* e *y*. (TRIOLA, 1999).

Na regressão múltipla, o objetivo é elaborar um modelo probabilístico que relacione uma variável dependente *y* a mais de uma variável independente ou de previsão (DEVORE, 1006).

Segundo Devore (2006), a equação do modelo de regressão múltipla é $y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$, até que se tenham colocados todos os fatores independentes.

O gráfico da equação de regressão, simples ou múltipla, é chamado de reta de regressão, ou reta de melhor ajuste. Esta definição expressa uma relação entre x (variável independente ou preditora) e y (variável dependente ou resposta) (JAMES *et al*., 2013). Na equação a representa o intercepto *y,* e b é o coeficiente angular da reta resultante.

O teste de regressão pode ser aplicado tanto da forma *enter* quanto *stepwise*. No modelo *enter*, todas as variáveis presentes no modelo são incluídas no resultado final, enquanto no método *stepwise*, as variáveis que se determinarem com pouca influência à variável dependente, são excluídas da solução final (FERREIRA, 2013).

Os resultados obtidos a partir do teste de correlação são a porcentagem de acerto do modelo e o quanto as variáveis independentes influenciam a variável dependente. Além disso, a solução final apresenta os coeficientes da equação de regressão, para que se possa obter uma previsão para futuros dados.

METODOLOGIA

Esta seção tem como objetivo apresentar como foram selecionados os dados, como foram construídas as tabelas de resultados e quais critérios foram adotados para alcançar o objetivo geral do trabalho.

Como mencionado anteriormente, este trabalho busca usar uma série histórica ao longo dos anos da média da expectativa de vida (em anos), e a população total brasileira (em quantidade nascimento de pessoas), além da série histórica do Produto Interno Bruto do Brasil (PIB). A série histórica vai de 1960 até 2016. O PIB é dado em quantidade de reais (R$) do total produzido pelo país (THE WORLD BANK, 2018). Na Tabela 1 consta os dados coletados da pesquisa.

Tabela 1 - Série histórica das variáveis PIB, expectativa de vida e população do Brasil

Anos	PIB, (US$ atual)	Expectativa de vida no nascimento, total (anos)	População, total
1960	$15.165.569.912,52	54,241	72.207.554
1961	$15.236.854.859,47	54,754	74.351.763
1962	$19.926.293.839,02	55,27	76.573.248
1963	$23.021.477.292,21	55,785	78.854.019
1964	$21.211.892.259,99	56,293	81.168.654
1965	$21.790.035.117,19	56,794	83.498.020
1966	$27.062.716.577,91	57,289	85.837.799
1967	$30.591.834.053,97	57,776	88.191.378
1968	$33.875.881.876,37	58,254	90.557.064
1969	$37.458.898.243,86	58,717	92.935.072
1970	$42.327.600.098,24	59,154	95.326.793
1971	$49.204.456.700,45	59,557	97.728.961
1972	$58.539.008.786,37	59,919	100.143.598
1973	$79.279.057.730,83	60,241	102.584.278
1974	$105.136.007.528,76	60,527	105.069.367
1975	$123.709.376.567,89	60,782	107.612.100
1976	$152.678.020.452,83	61,017	110.213.082
1977	$176.171.284.311,76	61,244	112.867.867
1978	$200.800.891.870,16	61,476	115.577.669
1979	$224.969.488.835,18	61,721	118.342.626
1980	$235.024.598.983,26	61,983	121.159.761
1981	$26.561.088.977,13	62,264	124.030.908
1982	$281.682.304.161,04	62,559	126.947.365
1983	$203.304.515.490,80	62,866	129.882.321
1984	$209.023.912.696,84	63,185	132.800.684
1985	$222.942.790.435,30	63,514	135.676.281
1986	$268.137.224.729,72	63,852	138.499.464
1987	$294.084.112.392,66	64,197	141.273.488
1988	$330.397.381.998,49	64,552	144.001.542
1989	$425.595.310.000,00	64,917	146.691.981
1990	$461.951.782.000,00	65,3	149.352.145
1991	$602.860.000.000,00	65,708	151.976.577
1992	$400.599.250.000,00	66,144	154.564.278
1993	$437.798.577.639,75	66,607	157.132.682
1994	$558.111.997.497,26	67,095	159.705.123
1995	$785.643.456.467,26	67,6	162.296.612
1996	$850.425.828.275,79	68,112	164.913.306
1997	$883.199.443.413,73	68,622	167.545.164
1998	$863.723.395.088,32	69,12	170.170.640
1999	$599.388.879.704,63	69,599	172.759.243
2000	$655.421.153.320,58	70,055	175.287.587
2001	$559.372.502.338,24	70,486	177.750.670
2002	$507.962.741.819,92	70,896	180.151.021
2003	$558.320.116.997,08	71,29	182.482.149
2004	$669.316.239.316,24	71,67	184.738.458
2005	$891.629.970.423,92	72,04	186.917.361
2006	$1.107.640.325.472,35	72,405	189.012.412
2007	$1.397.084.381.901,29	72,768	191.026.637
2008	$1.695.824.517.395,57	73,129	192.979.029
2009	$1.667.019.605.881,76	73,488	194.895.996
2010	$2.208.871.646.202,82	73,838	196.796.269

2011	$2.616.201.578.192,25	74,174	198.686.688
2012	$2.465.188.674.415,03	74,488	200.560.983
2013	$2.472.806.919.901,67	74,777	202.408.632
2014	$2.455.993.200.170,00	75,042	204.213.133
2015	$1.803.652.649.613,75	75,284	205.962.108
2016	$1.796.186.586.414,45	75,509	207.652.865

Para análise estatística dos dados, utilizou-se o *Software IBM SPSS Statistic*, o qual é líder de mercado para esse tipo de estudo. Esse recurso oferece uma quantidade ampla de relatórios, análises e gráficos, como teste de hipóteses (IBM, 2018). A utilização e licença desse software foram concedidos pela própria Universidade Tecnológica Federal do Paraná.

Considerando que os dados, segundo definição de variáveis, são um conjunto de variáveis quantitativas, estes foram organizados no software de maneira que atendessem suas próprias características, ou seja, a quantidade de casas decimais corretas, a nomenclatura como variáveis quantitativas, entre outros.

Após feita transferência dos dados para o software, foram realizadas as devidas análises para a discussão dos resultados. Além disso, também foi observado se seria possível fazer a correlação dos dados e a regressão linear múltipla da interferência das séries, expectativa de vida e população, sob o PIB brasileiro. O software oferece todas as ferramentas para a análise em discussão.

Foram, portanto, realizadas análises de estatística descritiva de cada série temporal, matriz de correlação, e o todos os testes necessários para observar se a regressão linear poderia ser aplicada, adotando o PIB como variável dependente e expectativa de vida e população como variáveis independentes. Após feita a simulação, foram coletados os principais resultados para análise.

RESULTADOS E DISCUSSÃO

Através da inserção dos dados no software foi possível obter os resultados para análise de estatística descritiva, testes de normalidade, correlação e regressão linear múltipla para as variáveis apresentadas.

ESTATÍSTICA DESCRITIVA

A partir da inserção dos dados de PIB, expectativa de vida e população no software, pôde-se observar, primeiramente, as estatísticas descritivas das variáveis, como média, mínimos e máximos, desvio padrão e número de dados na amostra. Os resultados obtidos para essa análise inicial estão expostos nos Quadros 1 e 2.

Quadro 1 - Estatística descritiva 1

Estatística Descritiva	PIB_US$ atual	Expec_Vida	População	N válido (de lista)
N	57	57	57	57
Mínimo	15.165.569.913	54,2410	72.207.554	
Máximo	2.616.201.578.192,25	75,5090	207.652.865	
Média	634.457.987.835,85	65,2622	142.325.306,6	
Desvio Padrão	73.577.9517.017,51610	6,2578	42.135.347,42	

Fonte: Adaptado IBM SPSS Statistic

Quadro 2 - Estatística descritiva 2

Estatística Descritiva		PIB_US$ atual	Expec_Vida	População	N válido (de lista)
Assimetria	Estatística	1,483	0,061	-0,07	
	Erro Padrão	0,316	0,316	0,316	
Curtose	Estatística	1,178	-1,173	-1,316	
	Erro Padrão	0,623	0,623	0,623	

Fonte: Adaptado IBM SPSS Statistic

Pode-se observar a partir das tabelas, que o número de dados em todas as variáveis foi de 57. Nota-se também, que a média do PIB, da expectativa de vida e da população, são, respectivamente, 634.457.987.835,85, 65,26221 e 142.325.306,6.

TESTE DE NORMALIDADE

A partir da inserção dos dados no software, também pode-se definir a natureza das variáveis com relação a normalidade. Como a quantidade de dados a serem analisados é superior a 30, utiliza-se o método Kolmogorov-Smirnov. Os resultados obtidos para essa análise estão situados no Quadro 3.

Quadro 3 – Testes de normalidade

Teste de Normalidade		PIB_US$ atual	Expec_Vida	População
Kolmogorov-Smirnov[a]	Estatística	0,201	0,079	0,081
	Df	57	57	57
	Sig.	0	0,200*	0,200*
Shapiro-Wilk	Estatística	0,779	0,954	0,941
	Df	57	57	57
	Sig.	0	0,029	0,007

Fonte: Adaptado IBM SPSS Statistic

A partir dos resultados obtidos, pode-se concluir que os dados de expectativa de vida e de população são considerados paramétricos, pois possuem significância superior a 0,05. Já os dados de PIB não possuem normalidade, pois sua significância obtida foi inferior a 0,05.

CORRELAÇÃO

A tabela de correlações foi obtida através do teste Correlação de Pearson, já que os dados de população e expectativa de vida são considerados paramétricos. Nesse teste, pode-se saber se as variáveis analisadas possuem correlação significativa e a força dessa correlação, conforme mostrado no Quadro 4.

Com base no valor de correlação e significância encontrados, pode-se concluir que população e expectativa de vida apresentam correlação significativa, pois *sig*. é menor que 0,05, e uma correlação forte, pois *r* é 0,995.

Quadro 4 – Correlações entre expectativa de vida e população

Correlações Expec_Vida X População		**Expec_Vida**	**População**
Expec_Vida	Correlação de Pearson	1	0,995**
	Sig. (2 extremidades)		0
	N	57	57
População	Correlação de Pearson	0,995**	1
	Sig. (2 extremidades)	0	
	N	57	57

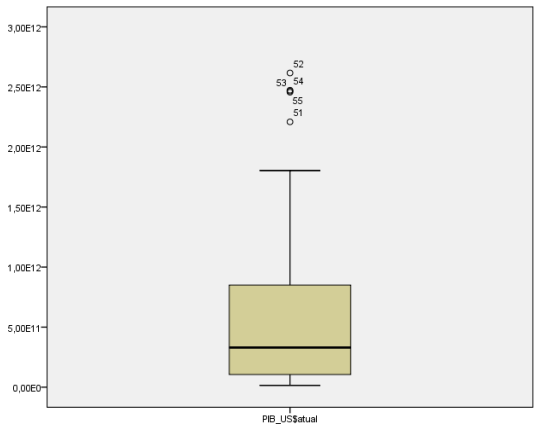
Fonte: Adaptado IBM SPSS Statistic

REGRESSÃO LINEAR MÚLTIPLA

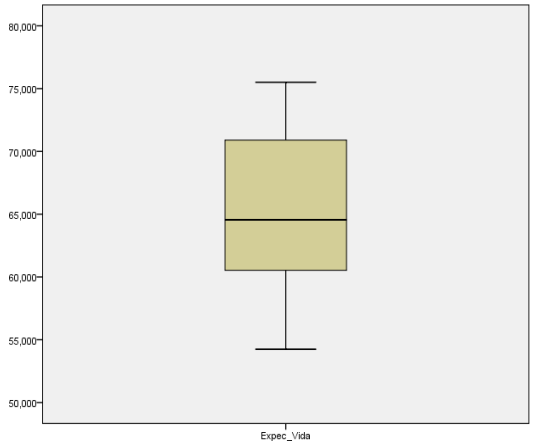
Como nem todas as variáveis em uso são consideradas paramétricas, pode ocorrer alguma pequena distorção nos resultados da regressão, mas a análise poderá ser satisfatória da mesma maneira.

Para a análise de regressão das variáveis, primeiramente deve-se garantir a pouca ou não existência de *outliers*. Esse teste foi feito através do *boxplot* presente na Figura 4.

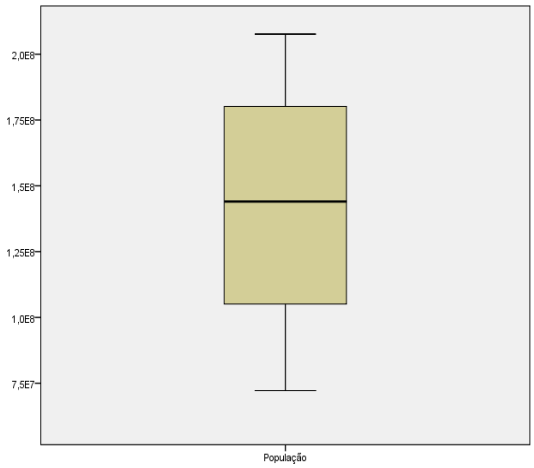
Figura 4 - Gráficos de dispersão (a) PIB, (b) Expectativa de vida e (c) população



(a)



(b)



(c)

Fonte: Adaptado IBM SPSS Statistic

Através da Figura 4(a), pode-se concluir que existem alguns *outliers* na variável PIB, mas, como há presença de muitos dados, há possibilidade desses pontos não interferirem na análise final. As variáveis expectativa de vida e população não possuem pontos extremos. O próximo passo para a afirmação de cumprimento de pressupostos é a análise de multicolinearidade das variáveis, como apresentado no Quadro 5.

Quadro 5 – Correlações

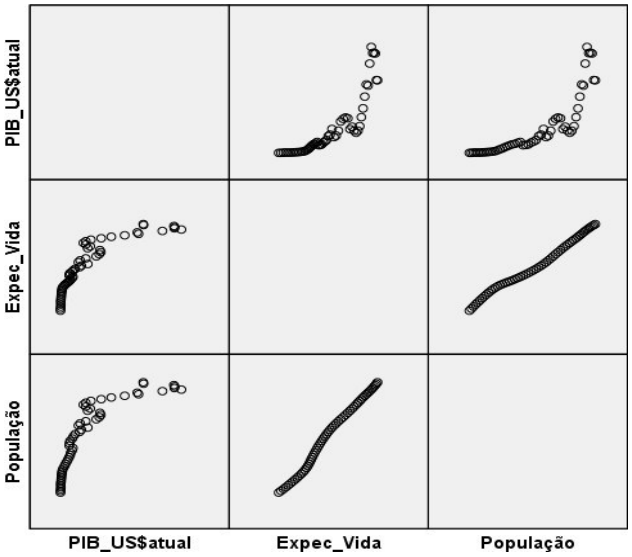
Correlações		PIB_US$ atual	Expec_Vida	População
Correlação de Pearson	PIB_US$atual	1	0,855	0,833
	Expec_Vida	0,855	1	0,995
	População	0,833	0,995	1
Sig. (1 extremidade)	PIB_US$atual	.	0	0
	Expec_Vida	0	.	0
	População	0	0	.
N	PIB_US$atual	57	57	57
	Expec_Vida	57	57	57
	População	57	57	57

Fonte: Adaptado IBM SPSS Statistic

Com base nos valores de correlação entre a expectativa de vida e população, PIB e população, PIB e expectativa de vida, pode-se afirmar que existe multicolinearidade entre expectativa de vida e população, pois a correlação é superior a 0,9. A partir disso, deveria-se eliminar uma das duas variáveis já que é sabido que haverá distorção no resultado final. Neste artigo, entretanto, a escolha foi de mantimento de todas as variáveis.

O último passo antes da realização da regressão é a avaliação da linearidade entre as variáveis dependentes e independentes através de gráficos de dispersão, como mostrado na Figura 5.

Figura 5 - Gráficos de dispersão



Fonte: Adaptado IBM SPSS Statistic

Com base nos gráficos de dispersão encontrados, pode-se observar que expectativa de vida e população apresentam uma linearidade crescente bem definida, enquanto, expectativa de vida e PIB e população e PIB apresentam uma linearidade positiva, mesmo que menos evidente.

Após o cumprimento dos pressupostos, inicia-se a regressão propriamente dita. Para isso é necessário dizer qual o percentual de acerto do modelo escolhido para análise. Para isso, verifica-se o Quadro 6.

Quadro 6 - Resumo do modelo

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa
1	0,878[a]	0,77	0,762	359.181.307.681

Fonte: Adaptado IBM SPSS Statistic

Observa-se que a variável R quadrado ajustado ou também chamado de coeficiente de determinação é igual a 0,762. Esse valor indica que o modelo consegue explicar os valores observados em um percentual de 76,20%.

A tabela ANOVA fornece alguns resultados os quais mostram que não há interferências significativas no estudo, ou seja, todos os pressupostos estudados antes da regressão em que houve dúvidas, foram agora confirmados como cumpridos. O Quadro 7 apresenta os resultados da ANOVA.

Quadro 7 – ANOVA

Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	2335019843256735000000000,000	2	1167509921628367500000000,000	90,497	0,000[b]
	Resíduo	6966605436534290000000000,000	54	12901121178767205000000,000		
	Total	3031680386910164000000000,000	56			

Fonte: Adaptado IBM SPSS Statistic

O valor de significância é 0,001 o que indica que o pressuposto de linearidade das variáveis não foi ferido e pode ser dado continuidade na regressão linear.

O Quadro 8, dado pela regressão linear, apresenta o valor da constante e dos coeficientes angulares de cada variável independente

Quadro 8 – Coeficientes

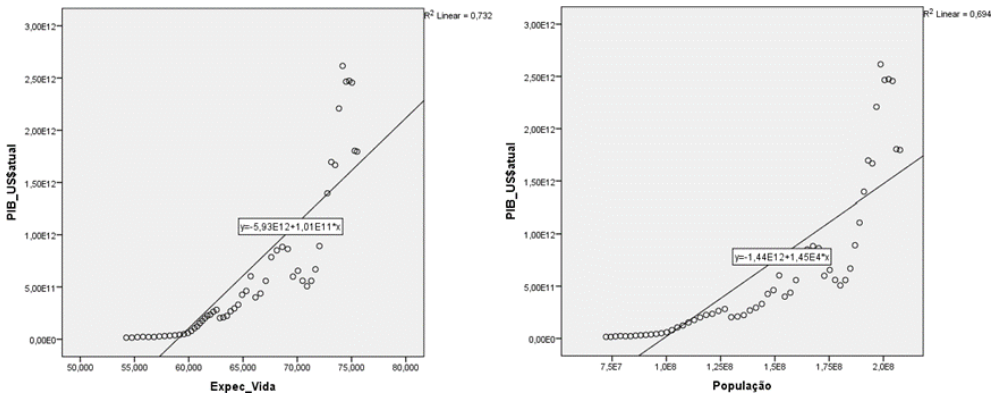
Modelo		Coeficientes não padronizados		Coeficientes padronizados	T	Sig.
		B	Erro Padrão	Beta		
1	(Constante)	- 16612276999707,37	3582555295010,3		-4,637	,000
	Expec_Vida	343214199882,383	80927230415,349	2,919	4,241	,000
	População	-36200,044	12019,096	-2,073	-3,012	,004

Fonte: Adaptado IBM SPSS Statistic

Com os valores fornecidos pela coluna Beta percebe-se que a variável expectativa de vida explica 291,90% e a variável População 207,30% da variável dependente em análise.

Os gráficos gerados a partir da reta de regressão para as variáveis estão presentes na Figura 6.

Figura 6 – Retas de regressão



Fonte: Adaptado IBM SPSS Statistic

A significância da constante e da variável expectativa de vida é igual a 0,001 enquanto da variável população é 0,004. As duas variáveis serão inseridas na equação da regressão linear e para montá-la analisa-se a coluna B do Quadro 8. A equação é dada a seguir:

$$y = -1{,}661x_1013 + 3{,}432x_1011x_1 - 36200{,}44x_2$$

onde:

y é a variável PIB.

x_1 é a variável Expectativa de vida x_2 é a variável População.

Os erros foram desconsiderados da equação. Para efeitos de análise foi realizada a previsão da variável PIB para os anos de 2017 a 2020, considerando que a base de dados utilizada possui dados até 2016. Os valores para as variáveis previsoras foram estimados baseados em estimativas dadas pelo Instituto Brasileiro de Geografia e Estatística e estão apresentados na Tabela 2.

Tabela 2 – Previsão das variáveis independentes

Variáveis previsoras	2017	2018	2019	2020
População	209251792,1	210779330,1	212254785,5	213676893
Expectativa de vida (anos)	75,99	76,25	76,5	76,74

Fonte: Autoria própria.

Logo em seguida os valores foram substituídos na equação para estimar o PIB e os valores estimados são apresentados na Tabela 3.

Tabela 3 – Previsão PIB

Previsão PIB (US$)			
2017	**2018**	**2019**	**2020**
$1{,}89476x10^{12}$	$1{,}9287x10^{12}$	$1{,}96108x10^{12}$	$1{,}99197x10^{12}$

Fonte: Autoria própria.

Realizada a previsão, conclui-se a regressão linear, com a observação de que os dados podem variar para mais ou para menos quando se considera o erro padrão.

CONCLUSÃO

O presente trabalho apresenta uma análise da correlação de 3 variáveis quantitativas e elaboração de um processo de regressão linear para realizar previsão do Produto Interno Bruto (PIB) do Brasil.

Ao realizar a análise das variáveis foi possível utilizar o método da regressão linear obtendo uma equação da reta para a previsão do PIB brasileiro para os anos de 2017 a 2020 baseado nas variáveis População Brasileira e Expectativa de vida.

Através da análise e discussão de resultados percebeu-se uma forte correlação positiva entre as variáveis, portanto à medida que a população e a expectativa de vida crescem, o PIB também cresce. Como já analisado, esse modelo proposto possui um percentual de acerto de 76,20% o que pode ser considerado aceitável para os dados e base escolhidos. Dessa maneira, considera-se os resultados da pesquisa como satisfatória, atendendo o objetivo geral de realizar uma análise estatística com a série histórica pesquisada.

Recomenda-se nesse estudo, como conclusão final, para que esse percentual de acerto do modelo aumente, devem ser inseridas mais variáveis que interfiram no PIB, como índice de desemprego nacional. Dessa forma, poderá se ter maior precisão da previsão da variável dependente.

Correlation and linear regression of variables that interfere in the Gross Domestic Product of Brazil: A data statistical analysis

ABSTRACT

This work aimed to analyze the correlation of 3 quantitative variables and to construct a linear regression equation to predict the dependent variable that is the Gross Domestic Product (GDP) of Brazil, using the independent variables: Life Expectancy and Population of Brazil. For this, data series were collected between 1960 and 2016 and, with the help of the IBM SPSS Statictic Software, we tried to analyze the interference of data from the independent variables on the dependent, through statistical analysis tools such as: Data normality tests, descriptive analysis of variables, correlation of variables and linear regression. As results, it was verified the possibility of using the linear regression method, obtaining a straight equation for the Brazilian GDP forecast for the years 2017 and 2020, using the proposed explanatory variables. This possibility was verified, since the population and life expectancy had positive and significant correlations in relation to GDP, and the percentage of correctness was 76.20%, which is considered acceptable for the selected databases. Thus, it was concluded that GDP grows or falls as changes in the explanatory variables of population and life expectancy occur.

KEYWORDS: Data statistical analysis. Correlation. Linear regression. Gross domestic product.

REFERÊNCIAS

BASTOS, E. K. X.; LAMEIRAS, M. A. P.; CARVALHO, L. M.; LEVY, P. M. **Economia brasileira no período 1987-2013:** relatos e interpretações da análise de conjuntura no Ipea. Brasília: Instituto de Pesquisa Econômica Aplicada (IPEA), 2015.

BURGEL, E.; ANZANELLO, M. J. Abordagem para seleção de variáveis preditivas no contexto de controle de inventários. **Revista Gestão Industrial**, Ponta Grossa, v. 14, n. 4, p. 154-195, out./dez. 2018. crossref

CAMARGOS, M. C. S; GONZAGA, M. R. **Viver mais e melhor? Estimativas de expectativa de vida saudável para a população brasileira**. 2015. Disponível em: http://www.scielo.br/pdf/csp/v31n7/0102-311X-csp-31-7-1460.pdf. Acesso em: 14 jun. 2018. crossref

CERVEIRA, D. S.; SELLITTO, M. A. Manutenção Centrada em Confiabilidade (MCC): análise quantitativa de um forno elétrico a indução. **Revista Produção Online**, Florianópolis, v.15, n. 2, p. 405- 432. abr./jun. 2015. crossref

DANCEY, C; REIDY, J**. Estatística Sem Matemática para Psicologia: Usando SPSS para Windows** 3ªEdição - Tradução Lorí Viali. Porto Alegre: Artmed., 2006.

DEVORE, J.L. **Probabilidade e Estatística: para Engenharia e Ciências** 6ªEdição. São Paulo: Pioneira Thomson Learning., 2016.

DIÃO, B.; DING, L.; SU, P.; CHENG, J. The spatial-temporal characteristics and influential factors of nox emissions in China: A spatial econometric analysis. **International Journal of Environmental Research and Public Health**. v. 15, n. 7,4. p. 1-19. Jul. 2018. crossref

FERREIRA, M. C. C. dos S. **Modelos de Regressão**: uma aplicação o em Medicina Dentária. 143f. 2013. Dissertação (Mestrado) – Universidade Aberta, Lisboa, 2013.

FIGUEIREDO, D. B. F; SILVA, J. A. J**. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r)***. 2009. Disponível em: <http://bibliotecadigital.tse.jus.br/xmlui/bitstream/handle/bdtse/2766/desvendando_mist%C3%A9rios_coeficiente_figueiredo%20filho.pdf?sequence=1>. Acesso em: 12 jun. 2018.

GOVBR – Governo do Brasil. Entenda como é medido o Produto Interno Bruto (PIB). 2016. Disponível em: <http://www.brasil.gov.br/economia-e-

emprego/2016/06/entenda-como-e-medido-o-produto-interno-bruto-pib> Acesso em: 07 jul. 2019.

HAIR, J. F. JR. et al. **Análise Multivariada de dados** 6ªEdição. São Paulo: Bookman Companhia Editora., 2009

IBGE. Projeção da população do Brasil e das Unidades Federativas. Disponível em: <https://www.ibge.gov.br/apps/populacao/projecao/> Acesso em: 12 jun. 2018.

IBM. **O que é IBM SPSS Statistics?**. Disponível em: <https://www.ibm.com/br- pt/marketplace/spss-statistics> Acesso em: 15 jun. 2018.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. New York: Springer, 2013. crossref

MANDAL, B.; BATINA, R.G.; CHEN, W. Do gender gaps in education and health affect economic growth? A cross-country study from 1975 to 2010. **Health Economics (United Kingdom)**. v. 27, n. 5, p. 877-886, may. 2018. crossref

MIOT, H. A. Avaliação da normalidade dos dados em estudos clínicos e experimentais. **Jornal Vascular Brasileiro**, Porto Alegre. v. 16, n. 2, p. 88-91, Abr./Jun. 2017. http://dx.doi.org/10.1590/1677-5449.041117. crossref

OLIVEIRA, G.; TUROLLA, F. Política econômica do segundo governo FHC: mudança em condições adversas. **Tempo social**. v. 15, n. 2, p. 195-217. 2003. http://dx.doi.org/10.1590/S0103-20702003000200008. crossref

PASSOS, A.G; MACIEL, M. A. C.; DORIA, M. R.; OLIVEIRA, R. B.; RUSSO, S. L. Análise estatística da evolução do produto interno bruto da indústria da construção civil brasileira utilizando regressão linear simples. **Revista Geintec**. n. 5, p. 505-514, 2012. crossref

PILATTI, L. E.; PICININ, C. T.; NASCIMENTO, R. F. O cenário da logística reversa em empresas multinacionais do município de Ponta Grossa-PR de 2010 e 2012. **Revista Gestão Industrial**, Ponta Grossa, v. 13, n. 1, p. 120-136, jan./mar. 2017. crossref

RATTANAMETAWEE, W.; LEENAWONG, C.; NETISOPAKUL, P. The effects of special events on regression for subcompact car sales in Thailand. **Jurnal Teknologi**. v. 78, n. 11, p. 161-165, nov. 2016.

RAZALI, N. M.; WAH, Y. B. Power comparisons of Shapiro-wilk, kolmogorovsmirnov, lilliefors and anderson-darling tests. **Journal of Statistical Modeling and Anlytics**. v. 2, p. 21-33. 2011.

Reis, E.A., Reis I.A. **Análise Descritiva de Dados.** Relatório Técnico do Departamento de Estatística da UFMG. 2002 Disponível em: <http://www.est.ufmg.br/portal/arquivos/rts/rte0202.pdf.> Acesso em: 12 jun. 2018.

RIBAS, J. R; VIEIRA, P. R. da C. **Análise Multivariada com o uso do SPSS**. Rio de Janeiro: Ciência Moderna Ltda., 2011.

RIBEIRO, F. C. S. et al. **A evolução do produto interno bruto brasileiro entre 1993 e 2009**. 2010. Disponível em: <http://img.fae.edu/galeria/getImage/1/1395677446523294.pdf.> Acesso em: 14 jun. 2018.

SILVA JÚNIOR, I. F.; OLIVEIRA, V. C. A aplicação do controle estatístico de processo numa indústria de beneficiamento de camarão marinho no estado do rio grande do Norte. **Revista Gestão Industrial**, Ponta Grossa, v. 01, n. 03, p. 59-69, 2005.

RICARDO, F. J.; MEDEIROS, L.; SALAS, C. S. S. Priorização de ações de eficiência energética para redes de hipermercados via análise multicritério. **Revista Gestão Industrial**, Ponta Grossa, v. 14, n. 1, p. 160-179, jan./mar. 2018.

THE WORLD BANK. Página Institucional. Disponível em: <https://data.worldbank.org/country/BR?locale=pt>. Acesso em: 10 jun. 2018.

TIBULO, C. et al. **Evolução populacional do Brasil: Uma visão demográfica**. 2012 Disponível em: <https://scientiaplena.emnuvens.com.br/sp/article/viewFile/772/447.> Acesso em: 14 jun 2018.

TORRES JÚNIOR, N.; NASCIMENTO, J. Z.; SOUZA, G. G. Análise de processos no ensino de graduação: uma estratégia didática baseada no uso conjunto da simulação computacional e do planejamento e análise de experimentos. **Revista Gestão Industrial**, Ponta Grossa, v. 09, n. 04, p. 930-952, 2013.

TRIOLA, M. F. **Introdução à Estatística** 7ªEdição. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora S.A., 1999.

VIANA, H. R. G.; MARQUES, A.; BIRANI, S.; SENA, S.; NOBUMASA, G. H. Manutenção centrada em confiabilidade: aplicação em motoredutores de transportadores de correias em uma refinaria de alumina. **Revista Gestão Industrial**, Ponta Grossa, v. 14, n. 2, p. 186-205, abr./jun. 2018.

Recebido: 11 abr. 2019
Aprovado: 09 jul. 2019
DOI: 10.3895/gi.v15n2.9968
Como citar:
KREMER, G. M.; DEINA, C.; SIQUEIRA, H. Correlação e regressão linear de variáveis que interferem no produto interno bruto do brasil: uma análise estatística de dados. R. Gest. Industr., Ponta Grossa, v. 15, n. 2, p. 233-254, abr./jun. 2019. Disponível em: <https://periodicos.utfpr.edu.br/rgi>. Acesso em: XXX.
Correspondência:
Guilherme Mateus Kremer
Av. Monterio Lobato, s/n, Jd. Carvalho, Ponta Grossa, Paraná, Brasil.