25º Seminário de Pesquisa do CCSA
As controvérsias da inovação e tecnologia no mundo do trabalho
11 a 15 de setembro de 2023

GT - MÉTODOS QUANTITATIVOS APLICADOS À ECONOMIA

# ALGORITMO GENÉTICO: uma aplicação em economia

Andre Gustavo Campos Pereira, Vagner dos Santos Torres, Janaina da Silva Alves

## RESUMO

Ao iniciar um estudo empírico, seja em economia ou em qualquer outra área do conhecimento, é comum ter um conjunto inicial muito grande de variáveis explicativas uma vez que não se sabe qual delas realmente contribui para explicar a variável resposta do problema. O objetivo do presente trabalho é apresentar uma ferramenta para auxiliar na escolha do subconjunto de variáveis explicativas que melhor ajusta o modelo de regressão analisado. Essa ferramenta utiliza o algoritmo genético elitista para encontrar o subconjunto de variáveis explicativas que minimizam o valor do AIC dos modelos analisados. Usamos um modelo macroeconômico do produto como função do consumo, investimento, gasto do governo, importações e exportações e acrescentamos outras variáveis para emular um desconhecimento das verdadeiras variáveis explicativas. Depois de escolhido o melhor conjunto de variáveis procedemos verificações a cerca de multicolinearidade, heterocedasticidade e autocorrelação no modelo estimado. O modelo escolhido pelo algoritmo apresentou alguns aspectos positivos para modelos que tem como objetivo apenas realizar previsões, não apresentou erro de especificação e o AIC foi significativamente menor que o modelo tradicional.

**Palavras-chave:** Regressão Linear, Algoritmo Genético, Akaike.

## ABSTRACT

When starting an empirical study, whether in economics or any other field of knowledge, it is common to have a very large initial set of explanatory variables since it is not known which ones truly contribute to explaining the response variable of the problem. The objective of this study is to present a tool to assist in selecting the subset of explanatory variables that best fits the analyzed regression model. This tool uses the elitist genetic algorithm to find the subset of explanatory variables that minimize the value of the AIC (Akaike Information Criterion) for the analyzed models. We used a macroeconomic model of output as a function of consumption, investment, government spending, exports, and added other variables to simulate a lack of knowledge about the true explanatory variables. After selecting the best set of variables, we conducted checks for multicollinearity, heteroscedasticity, and autocorrelation in the estimated model. The model chosen by the algorithm presented some positive aspects for models that aim solely at making predictions. It did not have specification errors, and the AIC was significantly lower than the traditional model.

**Keywords**: Linear Regression, Genetic Algorithm, Akaike

# 1 INTRODUÇÃO

Ao iniciar um estudo empírico, seja em economia ou em qualquer outra área do conhecimento, é comum ter um conjunto inicial muito grande de variáveis explicativas uma vez que não se sabe qual delas realmente contribui para explicar a variável resposta do problema. Muitas vezes temos uma teoria que respalda a escolha das variáveis que deve compor o modelo, outras vezes não. Como proceder nesse segundo caso?

Nos cursos de econometria, vemos que uma primeira abordagem é estimar o modelo via mínimos quadrados ordinários e nesse processo de estimação as variáveis estatisticamente significativas são obtidas[1].

Uma outra abordagem é via seleção de variáveis. O problema de seleção de variáveis tornou-se especialmente importante no contexto dos modelos lineares (RAO; TOUTENBURG, 1995) porque lidar com modelos com uma grande quantidade de variáveis pode ser teórica e computacionalmente difícil. Devido a esses fatores, surgiram muitas técnicas visando resolver esse problema no contexto linear. Porém, tais técnicas apresentam muitas restrições teóricas (como diferenciação ou convexidade) e práticas (como exigência de muitos recursos computacionais e instabilidade na convergência para uma solução) nos modelos. Assim, apresentamos neste trabalho uma técnica alternativa para seleção de variáveis que permite evitar tais restrições: o algoritmo genético.

Visando contornar os problemas citados acima, muitas técnicas foram desenvolvidas para resolver o problema de seleção de variáveis. Dentre as técnicas existentes podemos citar: RIDGE, análise de componentes principais, LASSO, seleção progressiva e regressiva de variáveis e seleção do melhor subconjunto de variáveis (RISH; GRABARNIK, 2014) (JAMES *et al*., 2013). Na Seção 2.2 mostramos como se dá a seleção do melhor subconjunto de variáveis e vemos que tal método requer um esforço computacional enorme caso o número de variáveis seja grande. Nessa abordagem o melhor modelo é aquele que apresenta o menor erro e podemos

[1] Vide (GREENE,2011), entre outros.

usar várias formas de se medir esse erro, dentre elas podemos citar o critério de informação de Akaike (AIC), o critério de informação Bayesiana (BIC) ou o Hannan-Quinn (HQ), etc. Essas abordagens se utilizam da verossimilhança do modelo como é explicado na Subseção 2.3.

Uma forma alternativa de se resolver o problema de seleção de variáveis sem ter que testar todas as possíveis soluções é usando os algoritmos estocásticos. Algoritmos estocásticos são aqueles que usam métodos probabilísticos em sua evolução. O algoritmo estocástico usado neste trabalho é o algoritmo genético elitista. O Algoritmo Genético apresentado em (HOLLAND, 1975), é uma ferramenta computacional que tenta emular o processo evolucionário de Darwin, o qual se utiliza de três estágios, a saber: Seleção, Cruzamento e Mutação. Esse tipo de algoritmo é usado para encontrar o ponto de ótimo aproximado de uma dada função $f: A \rightarrow \mathbb{R}$, ver (GOLDBERG, 1989).

Neste trabalho explicamos como usar o algoritmo genético elitista para encontrar o conjunto de variáveis que tenha o menor AIC dentre todos os modelos considerados. Essa ideia já foi usada em vários tipos de problemas envolvendo modelos de regressão, ver (ACOSTA-GONZÁLEZ; FERNÁNDEZ-RODRÍGUEZ, 2007; LACERDA; CARVALHO; LUDERMIR, 2002; PATERLINI; MINERVA, 2010). Aplicamos essa técnica num modelo de Macroeconomia que relaciona o Produto Interno Bruto - PIB de um país com o consumo das famílias, investimento, gastos do governo, importações e exportações. A fim de emularmos a falta de conhecimento de uma teoria que modela o problema, acrescentamos o quadrado, o inverso e o log de cada uma das variáveis do modelo original, fornecendo dessa forma um modelo com muitas variáveis explicativas. Os dados são trimestrais e foram obtidos no site do Instituto de Geografia e Estatística - IBGE, engloba o período compreendido do terceiro trimestre de 2000 até o terceiro trimestre de 2022 e foram deflacionados pelo IPCA do terceiro trimestre de 2000. As variáveis utilizadas são: PIB (P), Consumo (C), investimento (I), gasto do governo (G), exportações (X), importações (M) além do quadrado, do inverso e do log dessas mesmas variáveis.

Este trabalho é dividido em cinco seções. Na Seção 2 relembra-se a teoria dos algoritmos genéticos e descreve-se a versão que é utilizada neste trabalho, na Seção 3, modela-se o problema de modo a satisfazer as hipóteses do algoritmo genético, na Seção 4 as simulações numéricas são descritas e os resultados numéricos analisados. Por fim, na Seção 5 são apresentadas as considerações finais.

## 2 METODOLOGIA

Para construir a ferramenta objeto deste trabalho precisamos: gerar subconjuntos do conjunto de variáveis explicativas do modelo de regressão em estudo, mensurar a qualidade do ajuste do modelo que tenha apenas as variáveis escolhidas e minimizar essa medida em relação a todos os conjuntos possíveis de variáveis escolhidas. A primeira e terceira etapas são feitas pelo algoritmo genético elitista (AGE) e a segunda pelo critério de informação de Akaike (AIC). Na primeira etapa cada seleção de variáveis pode também ser vista como a seleção de um modelo. Explicamos a seguir como cada um desses itens funcionam isoladamente e na próxima seção como juntamos todos eles para trabalharem na busca do modelo procurado.

### 2.1 Algoritmo Genético Elitista - AGE

O Algoritmo Genético apresentado em (HOLLAND, 1975), é uma ferramenta computacional que tenta emular o processo evolucionário de Darwin, o qual se utiliza de três estágios, a saber: Seleção, Cruzamento e Mutação. Esse tipo de algoritmo é usado para encontrar o ponto de ótimo aproximado de uma dada função $f: A \to \mathbb{R}$, que é chamada função objetivo.

Para executar os passos do algoritmo o conjunto $A$ deve ser discretizado, ou seja, constrói-se um subconjunto $D \subset A$ de modo que cada ponto seja representado por vetores binários de comprimento l, onde l depende da precisão desejada. Sem prejuízo para o entendimento, como cada ponto de $D$ é identificado como um vetor binário, pode-se imaginar que os pontos de $D$ serão esses vetores binários. Precisa-

se também de populações com $N$ indivíduos e $Z = \{(u_1, u_2, \dots, u_N); u_i \in D, i = 1,2, \dots, N\}$ é o conjunto de tais populações e cada $u_i$ é um vetor binário de comprimento l. Na teoria dos algoritmos genéticos, cada ponto de $D$ é chamado de uma possível solução do problema de maximização/minimização da função $f$, $D$ também é chamado espaço de busca.

Em palavras, o algoritmo genético começa sua trajetória com um conjunto formado de $N$ possíveis soluções do problema e tenta, através dos operadores seleção, cruzamento e mutação, obter uma nova população que esteja mais próxima do ponto desejado como explicamos a seguir.

Cada passo do algoritmo pode ser resumido como a seguir:

- O operador seleção usa a definição da função $f$ para colocar dentro da população cópias dos pontos de maior imagem (no caso de maximização) ou de menor imagem (no caso de minimização) descartando os pontos com menores/maiores imagens dependendo do caso analisado.
- O operador cruzamento usa pares (o número de tais pares é definido pelo parâmetro probabilidade de cruzamento Pc) para gerar novos pares de possíveis pontos de ótimo. Essa etapa tenta emular a melhora genética dos pontos obtida via cruzamento.
- O operador de mutação muda os pontos da população através de um parâmetro (probabilidade de mutação Pm) tentando emular as mutações que o meio impõe aos membros da população.

O algoritmo genético foi projetado para

- "Melhorar" as populações seguintes a partir da população inicial.
- A "melhoria" da população seria ter os valores das imagens dos pontos dessa população pela função $f$ maiores (no caso de maximização) ou menores (no caso de minimização) do que as imagens dos pontos da população anterior.
- A ideia do algoritmo genético era que ele convergisse para uma população que fosse composta por cópias do ponto de ótimo procurado.

- Foi demonstrado em (RUDOLPH, 1994) que isso não acontece quase certamente (ou seja, com probabilidade 1).

Rudolph, no mesmo artigo (RUDOLPH, 1994) apresenta o algoritmo genético elitista (AGE) que resolve o problema de convergência mencionado anteriormente. Esse novo algoritmo evolui da seguinte maneira

a) Escolha aleatoriamente uma população inicial tendo $N$ elementos, cada um sendo um vetor binário de comprimento *l*, e crie mais uma posição, a $(N+1)$-ésima entrada do vetor população, a qual manterá o "melhor" elemento daqueles $N$ elementos anteriores.

b) Repita

1. Execute a seleção com os primeiros $N$ elementos;
2. Execute o cruzamento com os $N$ primeiros elementos;
3. Execute a mutação com os $N$ primeiros elementos;
4. Se o melhor elemento dessa nova população é melhor que aquele que está na $(N+1)$-ésima posição, troque a $(N+1)$-ésima posição por esse melhor elemento, caso contrário preserve a $(N+1)$-ésima posição inalterada.

c) Até que algum critério de parada seja atingido.

A única diferença do algoritmo genético para o algoritmo genético elitista é que o AGE vai levando o melhor ponto encontrado a medida que o algoritmo vai evoluindo, ou seja, o AGE não perde o melhor ponto encontrado ao longo de sua trajetória enquanto o algoritmo genético pode perder. Algumas versões convergentes desse algoritmo em que os parâmetros variam podem ser vistos em (PEREIRA *et al*., 2018; PEREIRA *et al*., 2019).

## 2.2 Seleção de variáveis

Seja $S = \{X_1, X_2, \dots, X_d\}$ um conjunto de variáveis aleatórias reais independentes, uma variável real $Y$ e um modelo de regressão linear múltipla

$$Y = h(\tilde{X}, \alpha) + \varepsilon$$

em que $\tilde{X} = (X_1, X_2, \dots, X_d)$ é um vetor de variáveis, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$ é o vetor de parâmetros do modelo (os coeficientes de regressão) e $\varepsilon$ é a variável aleatória

representando o erro do modelo $Y$. O problema de seleção de variáveis consiste em determinar um subconjunto $V \subset S$ de variáveis a serem incluídas no modelo. Equivalentemente, o problema de seleção de variáveis consiste em determinar quais são as componentes não-nulas do vetor de parâmetros $\alpha$.

Tal problema possui diversas abordagens e dentre elas a conhecida como: seleção do melhor subconjunto de variáveis. Tal método consiste em construir todos os modelos possíveis para a variável resposta $Y$ com as $d$ variáveis de $\tilde{X} = (X_1, X_2, \dots, X_d)$ e analisar qual modelo possui o menor nível de erro (quando considera-se algum critério para mensurar os erros cometidos nos dados)

$$\min_h erro(h)$$

onde $erro(h)$ é o tipo de função erro selecionada pelo usuário para mensurar numericamente o erro cometido pelo modelo $h$ nos dados disponíveis. No contexto da análise de regressão linear múltipla, frequentemente considera-se a função $h$ como

$$h(\tilde{X}, \alpha) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_d X_d.$$

Como as $d$ variáveis em $\tilde{X}$ permitem construir no máximo $2^d$ modelos com configurações distintas de variáveis, então deve-se realizar os seguintes procedimentos gerais para selecionar qual dos $2^d$ modelos, melhor se ajusta aos dados determinados por $\tilde{X} = (X_1, X_2, \dots, X_d)$ :

1. Inicia-se com um modelo sem nenhuma variável $h(0) = \alpha_0$ definido como um modelo constante e igual ao intercepto $\alpha_0$. Tal modelo está em um conjunto denotado por $M_0$.
2. Para cada $k \in \{1,2, \dots, d\}$ denota-se por $M_k$ o conjunto formado por todos os modelos que contêm exatamente $k$ variáveis. Pela identidade combinatorial

$$\sum_{j=0}^{d} \binom{d}{j} = 2^d,$$

3. a união dos conjuntos $M_k$ gera todos os modelos distintos possíveis com no máximo $d$ variáveis. Em seguida, escolhe-se em cada conjunto $M_k$ para $k \in \{1,2, \dots, d\}$ um modelo $h_k$ tal que

$$erro(h_k) = \min_{h \in M_k} erro(h)$$

4. gerando ao final um conjunto $M = \{h_1, \dots, h_d\}$.
5. A partir do conjunto $M$ escolhe-se o modelo com o melhor subconjunto de variáveis de $\{X_1, X_2, \dots, X_d\}$ como sendo o modelo $h_{k_0}$ de índice $k_0$ tal que

$$erro(h_{k_0}) = \min_{h \in M} erro(h).$$

Existem ainda outros métodos de seleção de variáveis que não serão discutidos no âmbito deste trabalho.

Quando se considera o modelo linear

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_d X_d + \varepsilon$$

e um conjunto de observações $(Y, X_1, X_2, \dots, X_d)$ pode-se reescrever o modelo no formato de produto matricial da seguinte forma

$$Y = \tilde{X}\alpha + \varepsilon$$

em que $\tilde{X} \in M_{m \times (d+1)}(\mathbb{R})$ é a matriz reunindo as observações das variáveis do modelo e $\varepsilon$ é o vetor aleatório associado ao modelo em questão.

Determinar o modelo linear acima é equivalente a estimar os parâmetros em $\alpha$ e o erro do modelo $\varepsilon$. Usando o método de mínimos quadrados, obtém-se tais estimativas dos parâmetros minimizando a função objetivo

$$h(\alpha) = (y - \tilde{X}\alpha)^t (y - \tilde{X}\alpha),$$

onde $t$ indica a transposta da matriz $y - \tilde{X}\alpha$. Em (RAO; TOUTENBURG, 1995) observa-se que encontrar uma solução para tal problema é equivalente a encontrar uma solução para o sistema linear

$$\tilde{X}^t \tilde{X}\alpha = \tilde{X}^t y$$

que no caso particular da matriz $\tilde{X}^t\tilde{X}$ ser inversível, tem-se que

$$\alpha = (\tilde{X}^t\tilde{X})^{-1}\tilde{X}^t y$$

é a solução do sistema linear acima.

No exemplo utilizado nesse trabalho, é levado em consideração a matriz das variáveis explicativas de posto completo e, portanto, os coeficientes do modelo ajustado são obtidos via a Equação (1).

2.3 O critério da informação de Akaike

O critério da informação de Akaike, popularmente conhecido como AIC (Akaike information criterion) é um estimador de uma medida numérica conhecida como informação de Kullback-Leibler (ANDERSON; BURNHAM; ANDERSON, 1998). Dadas duas funções $f$ e $g$, a informação de Kullback-Leibler é definida por

$$\boldsymbol{I(i,g) = \int_{\Omega} f(x) ln\left(\frac{f(x)}{g(x,\theta)}\right) dx.}$$

A importância de tal medida está no fato que dado um conjunto de dados gerado por um modelo $f$, a informação de Kullback-Leibler entre $f$ e $g$ determina a informação perdida ao usar o modelo $g$ para ajustar os dados gerados por $f$. Conhecendo a definição da informação de Kullback-Leibler pode-se derivar o AIC como um estimador de tal medida, cuja expressão é

$$AIC(g) = -2\ln\left(L(\hat{\theta})\right) + 2K,$$

em que $\hat{\theta}$ é o estimador de máxima verossimilhança dos parâmetros do modelo $g$ e $K$ é o número de parâmetros do modelo $g$. O AIC possui ampla aplicação para solucionar o problema de seleção de modelos, ou seja, pode-se buscar um modelo em um conjunto finito de modelos $M = \{g_1, g_2, \dots, g_m\}$ que melhor se ajusta aos dados gerados por um modelo desconhecido $f$. A seleção de modelos ocorre buscando qual modelo em $M$ possui menor valor de AIC e consequentemente melhor aproxima o modelo desconhecido $f$.

Especificamente quando tratando de modelos de regressão linear com erro distribuído normalmente, o AIC pode ser escrito como

$$AIC_{LM}(g) = nln\left(\frac{SQR(g)}{n}\right) + 2K,$$

em que $SQR(g)$ é a soma dos resíduos quadráticos do modelo $g$ usado para aproximar o modelo desconhecido $f$, $K$ o número de parâmetros do modelo $g$ e $n$ o número de observações disponíveis nos dados. Quando se considera a estimação dos parâmetros por meio do método dos mínimos quadrados, o número de

parâmetros estimados já inclui na contagem o intercepto e a variância do modelo, o que nos leva a observar o AIC como:

$$AIC_{LM}(g) = nln(\hat{\sigma}^2) + 2K,$$

em que

$$\hat{\sigma}^2 = \frac{SQR(g)}{n}.$$

Devido a estudos posteriores e limitações observadas na expressão inicial do AIC, estimulou-se novas buscas por estimadores da informação de Kullback-Leibler adequados para diferentes contextos dos dados. Nos trabalhos (HURVICH; TSAI, 1991) e (SUGIURA, 1978), perceberam que o AIC pode gerar estimativas imprecisas da informação de Kullback-Leibler se o número de parâmetros é grande em comparação a quantidade de dados disponíveis. Assim, derivaram um AIC ajustado a pequenas amostras chamado $AIC_c$ cuja expressão é:

$$AIC_c(g) = AIC(g) + \left(\frac{2K(K+1)}{n-K-1}\right),$$

onde $n$ é o tamanho da amostra disponível e $K$ o número de parâmetros do modelo $g$.

Apesar de existirem variações do AIC, todas elas possuem termos de penalidade dependendo de $K$ (o número de parâmetros do modelo). A existência de tais penalidades implica que quanto maior o número de parâmetros $K$ do modelo, maior o valor de AIC associado ao modelo. Como o objetivo da seleção de modelos é encontrar um modelo com o menor valor de AIC em $M$, pode-se simultaneamente realizar a seleção de variáveis a serem incluídas no modelo visto que dados dois modelos em $M$ com o mesmo valor de AIC o princípio da parcimônia (variante do princípio da navalha de Occam) declara que o modelo com a menor quantidade de parâmetros é mais adequado para aproximar o modelo desconhecido $f$. Dessa forma, obter um modelo com menos parâmetros não-nulos é equivalente a obter um modelo com menos variáveis (e suas transformações).

## 3 PROCEDIMENTOS METODOLÓGICOS

### 3.1 Modelando o problema

Suponha que exista uma variável resposta $Y$ e que não se sabe que tipo de relação ela tem com as variáveis explicativas $X_1, X_2, \dots, X_n$. Para simplificar o modelo imagine que existam duas variáveis explicativas $X_1$e $X_2$. Mesmo nessa situação simples podemos nos deparar com várias possibilidades para o modelo, por exemplo:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon$$

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 \ln(X_1) + \alpha_3 X_2 + \varepsilon$$

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 \ln(X_1) + \alpha_3 \frac{1}{X_1} + \alpha_4 X_2 + \alpha_5 \ln(X_2) + \alpha_6 \frac{1}{X_2} + \varepsilon$$

Como os modelos têm a mesma variável resposta, podemos montar o espaço de busca com todas as variáveis explicativas e depois rodar um AGE utilizando a função AIC como função objetivo, para decidir que variáveis devem entrar no modelo. O AGE vai evoluindo de forma que o melhor modelo ajustado (aquele com menor AIC) vai tendo uma cópia guardada durante o processo. Uma vez decidido o melhor modelo parte-se para todas as outras análises que fazem com que um modelo seja bom, a saber: Normalidade dos erros, não-especificidade etc.

### 3.2 Etapas

De posse dos dados das variáveis explicativas e da variável resposta, monta-se cada uma das variáveis explicativas, $X, \ln(X), \frac{1}{X}$. Nos exemplos apresentados nas equações (2) - (4), observa-se que o espaço utilizado pelo algoritmo genético terá sete entradas, uma para cada parâmetro, ou seja, a discretização do espaço de busca será composta por vetores binários de dimensão 7, a saber: $D = \{(a_1, a_2, \dots, a_7)/a_i \in \{0,1\}, i = 1,2, \dots, 7\}$.

Depois de estabelecido o espaço de busca, precisa-se definir o tamanho da população utilizada. Lembrando que existe um *trade-off* entre tamanho da população e a velocidade do algoritmo, quanto maior a população mais pontos o

algoritmo analisa, porém mais demorado será a execução de um passo do algoritmo. Assim, escolhido o tamanho $N$ da população, obtém-se o conjunto de todas as possíveis populações $Z = \{(u_1, u_2, \dots, u_N)/u_i \in D, i = 1, \dots, N\}$.

Falta agora estabelecer quem é a função objetivo, que é a função que se busca maximinizar/minimizar dependendo do problema, nesse caso a função será o AIC do modelo escolhido e o objetivo é minimizar tal função.

Vemos na descrição do AGE que existem dois parâmetros que devem ser fornecidos pelo usuário, a saber: A probabilidade de cruzamento ($p_c$) e a probabilidade de mutação ($p_m$). Uma vez definidas essas probabilidades temos todas as condições para rodar o algoritmo. Em geral, a probabilidade de mutação é pequena e a probabilidade de cruzamento é em torno de 0,5, ver (GOLDBERG, 1989).

Todas as exigências para se rodar o algoritmo foram atendidas, a saber: Um espaço de busca bem definido (conjunto de todos os possíveis modelos), função objetivo (AIC do modelo escolhido) e a definição dos parâmetros. A cada população, o modelo é ajustado de acordo com as variáveis escolhidas, o AIC de cada um deles é calculado e o AGE vai gerar novas populações, ou novos modelos que são os novos candidatos ao modelo com o menor AIC, ou seja, ao modelo mais ajustado. Depois de uma quantidade de passos pré-estabelecida, o algoritmo é parado e é escolhido o melhor modelo encontrado até aquele momento.

## 4 RESULTADOS

### 4.1 Simulações

Nessa seção a ferramenta desenvolvida é aplicada ao modelo macroeconômico simplificado, que relaciona o Produto Interno Bruto - PIB de um país com o consumo das famílias, investimento em bens de capital, gastos do governo, importações e exportações, a partir das contas nacionais (LOPES; VASCONCELLOS, 2008). Os dados coletados são trimestrais e foram obtidos no site do Instituto de Geografia e Estatística - IBGE, englobam o período compreendido do terceiro trimestre de 2000 até o terceiro trimestre de 2022 e foram deflacionados

pelo IPCA do terceiro trimestre de 2000. Sabe-se que o modelo tradicional é descrito de forma simplificada como:

$$P = \beta_0 + \beta_1 C + \beta_2 I + \beta_3 G + \beta_4 X + \beta_5 M + \varepsilon$$

Conforme (LOPES; VASCONCELLOS, 2008), os sinais dos parâmetros, na equação (5), são positivos para os coeficientes $\beta_1, \beta_2, \beta_3$ e $\beta_4$, tendo sinal negativo para o para parâmetro $\beta_5$. Já o parâmetro $\beta_0$, não tem significado econômico definido na literatura econômica.

As variáveis utilizadas para as simulações são: PIB (P), Consumo das famílias (C), investimento em bens de capital (I), gasto do governo (G), exportações (X) e importações (M) bem como seus quadrados, inversos e seus logaritmos.

Neste trabalho incluímos novas variáveis a fim de expressar o desconhecimento do usuário sobre a teoria que explica o modelo:

$$\begin{aligned} P = \beta_0 + \beta_1 C + \beta_2 C^2 + \beta_3 \frac{1}{C} + \beta_4 \ln(C) + \beta_5 I + \beta_6 I^2 + \beta_7 \frac{1}{I} + \beta_8 \ln(I) + \beta_9 G + \beta_{10} G^2 \\ + \beta_{11} \frac{1}{G} + \beta_{12} \ln(G) + \beta_{13} X + \beta_{14} X^2 + \beta_{15} \frac{1}{X} + \beta_{16} \ln(X) + \beta_{17} M \\ + \beta_{18} M^2 + \beta_{19} \frac{1}{M} + \beta_{20} \ln(M) + \varepsilon \end{aligned}$$

Uma vez estabelecidas as variáveis consideradas relevantes no modelo, roda-se o algoritmo e são executadas 5000 iterações obtendo-se a seguinte resposta como sendo o modelo que produziu o menor AIC (1980.991), dentre os modelos analisados pelo algoritmo.

$$\begin{aligned} P = 1{,}031 \times 10^6 + 1{,}338C + 1{,}132 \times 10^{-6} C^2 + 0{,}811I + 0{,}9971G - 1{,}524 \times 10^{-6} G^2 \\ + 1{,}32X - 6{,}54 \times 10^{-7} X^2 - 1{,}061 \times 10^5 \ln(M) + \varepsilon \end{aligned}$$

Logo, pode-se constatar a forma funcional resultante que representa a relação do PIB com as respectivas variáveis explicativas, já definidas anteriormente.

4.2 Resultados numéricos

Apresentam-se na Tabela 1 os resultados dos testes de significância estatística individual e conjunta do modelo, como também o valor do coeficiente de determinação ($R^2$) do modelo apontado pelo algoritmo. Os resultados foram obtidos no programa RStudio versão 2022.07.1.

Tabela 1: Modelo Escolhido pelo Algoritmo Genético

| | Estimate | Std. Error | t-value | Pr( >\| t \| ) | |
|---|---|---|---|---|---|
| Intercept | $1{,}031 \times 10^{6}$ | 296195.4351 | 3.48 | 0.0008 | *** |
| $C$ | 1,338 | 0.1186 | 11.28 | 0.0000 | *** |
| $C^2$ | $1{,}132 \times 10^{-6}$ | 0.0000 | -2.02 | 0.0467 | * |
| $I$ | 0.8110 | 0.1140 | 7.11 | 0.0000 | *** |
| $G$ | 0.9971 | 0.3176 | 3.14 | 0.0024 | ** |
| $G^2$ | $-1{,}524 \times 10^{-6}$ | 0.0000 | -3.22 | 0.0018 | ** |
| $X$ | 1.3197 | 0.2320 | 5.69 | 0.0000 | *** |
| $X^2$ | $-6{,}54 \times 10^{-7}$ | 0.0000 | -1.90 | 0.0617 | . |
| $\ln(M)$ | $-\ 1{,}061 \times 10^{5}$ | 29051.9738 | -3.65 | 0.0005 | *** |

Fonte: Elaboração própria.

Temos também o número de graus de liberdade de 80, um erro de padrão residual de 15550, o $R^2$ foi 0,9994 e $R^2$-ajustado foi 0.9993. O teste F(8,80) de significância foi $1{,}661 \times 10^{4}$ e o p-valor do teste F foi menor que $2{,}2 \times 10^{-16}$.

O modelo selecionado apresenta significância estatística individual para todas as variáveis ao nível de 5%. Exceto para X2, que apresentou significância estatística ao nível de 10%. O poder de explicação conjunta do modelo é significativo, como apontado pelo resultado do p-valor do teste F, muito próximo de zero. De acordo com o coeficiente de determinação $R^2$, o modelo explica 99,94% das variações da variável resposta. Os sinais dos parâmetros estimados apresentam coerência com os apontados por LOPES e VASCONCELLOS (2008). As variáveis C, I, G e X apresentam sinal positivo indicando impacto direto sobre o PIB. Já a variável M apresenta sinal negativo, indicando um impacto inverso das importações sobre o PIB. Ainda temos 0,42607 como o resultado do teste RESET com df1 = 2 e df2 = 78 e o p-valor desse teste foi de 0,6546.

As variáveis $C^2$ e $X^2$ apresentam sinal negativo, conforme apontado pelo algoritmo genético, mostrando que embora o consumo e as exportações contribuam

positivamente, a taxa de tal contribuição é decrescente (taxa decrescente de aumento, ou seja, a tendência é crescer cada vez menos).

Uma vez que o algoritmo genético busca identificar a forma funcional ótima para o modelo de regressão, testa-se inicialmente a ocorrência de erro de especificação. Através dos testes de Reset de Ramsey não foi possível identificar a ocorrência de erro de especificação, logo, há evidências de que a forma funcional foi escolhida de forma adequada.

Conforme Gujarati (2011), não existem testes formais para a detecção da multicolinearidade, porém, alguns aspectos podem indicar a ocorrência da multicolinearidade, tais como parâmetros sem significância estatística individual e $R^2$ elevado. O que de fato pode-se observar a partir dos resultados apresentados na Tabela 1. Ainda de acordo com Gujarati (2011) pode-se calcular o Fator de Inflação da Variância - FIV, como uma forma de identificar qual o grau da multicolinearidade do modelo. Como apresentado na Tabela 2, observa-se que as variáveis explicativas apresentam valores considerados elevados para o FIV, indicando a ocorrência da multicolinearidade.

As implicações da multicolinearidade podem não ser tão prejudiciais, a depender do objetivo pretendido com o modelo estimado. Para efeito de previsão, a ocorrência da multicolinearidade não chega a ser um problema grave. A dificuldade, no caso de ocorrência da multicolinearidade, é separar os efeitos individuais das variáveis explicativas na variável resposta. Segundo (MAIA, 2017) a colinearidade pode tratar-se de uma característica ou deficiência própria dos dados e que não há muito a ser feito. Mesmo assim, os estimadores de MQO ainda possuem propriedades de melhores estimadores lineares não viesados.

Tabela 2: Análise da Multicolinearidade

| Variável | FIV |
|---|---|
| $C$ | 765.19480 |
| $C^2$ | 427.31845 |
| $I$ | 54.64167 |
| $G$ | 511.54259 |
| $G^2$ | 262.03522 |
| $X$ | 254.36988 |

| | |
|---|---|
| $X^2$ | 144.67655 |
| $\ln(M)$ | 144.82161 |

Fonte: Elaboração própria

Quanto a presença de heterocedasticidade no modelo, onde ocorre a quebra do pressuposto da variação constante do termo de erro. Pode-se realizar testes formais para a detecção. Então recomenda-se, por exemplo, a realização dos testes de Breusch-Pagan e White, ambos detectaram a ocorrência de heterocedasticidade dos erros no modelo. Embora permaneçam não viesados e consistentes, os estimadores de MQO deixam de ser eficientes, ou seja, não mais apresentam variância mínima.

Porém, ignorar a existência desse problema e seguir com a estimação, pode-se chegar a conclusões e inferências equivocadas. Recomenda-se a tomada de algumas medidas corretivas. Uma vez que o $\sigma^2$ é desconhecido, logo, indica-se a estimação do modelo de regressão utilizando o erro padrão robusto de White. A Tabela 3 apresenta a estimação desses parâmetros. Como a heterocedasticidade pode ser causada por vários fatores e os estimadores consistentes de White corrigem apenas a heterocedasticidade, deve-se avaliar a ocorrência também de autocorrelação, (GUJARATI, 2011).

Tabela 3: Modelo de Regressão com erros Robustos de White

| Coefficients | Value | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 637455. | 193097.4849 | 3.3012 |
| $C$ | 1.4770 | 0.0773 | 19.0964 |
| $C^2$ | 0.0000 | 0.0000 | -4.8588 |
| $I$ | 0.5388 | 0.0743 | 7.2478 |
| $G$ | 0.9116 | 0.2071 | 4.4020 |
| $G^2$ | 0.0000 | 0.0000 | -4.2534 |
| $X$ | 0.7944 | 0.1512 | 5.2530 |
| $X^2$ | 0.0000 | 0.0000 | 1.0074 |
| $\ln(M)$ | -67983.4797 | 18939.7351 | -3.5895 |

Fonte: Elaboração própria

Temos ainda 80 graus de liberdade, um erro residual padrão de 9098.

Avalia-se a ocorrência de autocorrelação no modelo escolhido pelo algoritmo. Então aplicam-se os testes de Durbin-Watson e Breusch-Godfrey os quais indicam a ocorrência de autocorrelação. A estatística de Durbin-Watson

(1.3204) indica que há evidências da ocorrência de autocorrelação positiva no modelo.

Tabela 4: Modelo de Regressão com Erros CHA

| Coefficients | Value | Std. Error | t value | Pr( > \| t \|) | |
|---|---|---|---|---|---|
| (Intercept) | $1.0305 \times 10^{6}$ | $2.9620 \times 10^{5}$ | 3.4793 | 0.0008166 | *** |
| $C$ | 1.3383 | 0.11864 | 11.2804 | $<2.2 \times 10^{-16}$ | *** |
| $C^2$ | $-1.1323 \times 10^{-7}$ | $5.6039 \times 10^{-8}$ | -2.0205 | 0.0466767 | * |
| $I$ | 0.81096 | 0.11404 | 7.1112 | $4.335 \times 10^{-10}$ | *** |
| $G$ | 0.99708 | 0.31764 | 3.1391 | 0.0023741 | ** |
| $G^2$ | $-1.5239 \times 10^{-6}$ | $4.7284 \times 10^{-7}$ | -3.2228 | 0.0018375 | ** |
| $X$ | 1.3197 | 0.23198 | 5.6887 | $2.023 \times 10^{-7}$ | *** |
| $X^2$ | $2.023 \times 10^{-7}$ | $2.023 \times 10^{-7}$ | -1.8950 | 0.0617038 | . |
| $\ln(M)$ | $-1.0607 \times 10^{5}$ | $2.9052 \times 10^{4}$ | -3.6511 | 0.0004643 | *** |

Fonte: Elaboração própria

Conforme Gujarati (2011) recomenda-se utilizar o método cujos erros padrão são conhecidos como erros padrão Consistentes para Heterocedasticidade e Autocorrelação - CHA. Os resultados são apresentados na Tabela 4.

Por fim, estima-se o modelo na sua forma funcional original com o objetivo de realizar comparações entre os modelos. Conforme apresentado na Tabela 5. Todos os sinais dos parâmetros são coerentes com a literatura, sendo ambos positivos, com exceção da variável importações (M), quanto à significância estatística individual, com exceção dos gastos finais do governo (G) e do intercepto, todos os parâmetros foram significativos estatisticamente. O modelo apresenta ainda significância estatística global, logo o poder de explicação conjunta do modelo foi significativo, com coeficiente de determinação $R^2$ de 0,999, isso significa que 99,9% das variações no PIB são explicadas pelo conjunto de variáveis.

Tabela 5: Modelo de Keynesiano

| | Estimate | Std. Error | t value | Pr( > \| t \|) | |
|---|---|---|---|---|---|
| (Intercept) | 6409.9719 | 5481.3806 | 1.17 | 0.2456 | |
| $C$ | 1.2640 | 0.0461 | 27.44 | 0.0000 | *** |
| $I$ | 1.0994 | 0.0844 | 13.03 | 0.0000 | *** |
| $G$ | 0.1341 | 0.1144 | 1.17 | 0.2444 | |
| $X$ | 0.8875 | 0.1030 | 8.62 | 0.0000 | *** |
| $M$ | -1.0162 | 0.1549 | 6.56 | 0.0000 | *** |

Fonte: Elaboração própria.

Temos também o número de graus de liberdade de 83, um erro de padrão residual de 19540, o $R^2$ foi 0.999 e $R^2$-ajustado foi 0.999. O teste F(5,83) de significância foi $1{,}681 \times 10^4$ e o p-valor do teste F foi menor que $2{,}2 \times 10^{-16}$.

Igualmente analisa-se o valor do AIC (2054.196), em seguida analisa-se a ocorrência de multicolinearidade no modelo. Que foi constatada em menor grau em relação ao modelo escolhido pelo algoritmo, com valores do FIV entre 18.9201 e 72.99103. Quanto à especificação do modelo, os testes apontam para ocorrência de erro de especificação. Não se constata autocorrelação, como apontado pelos testes de Durbin-Watson e Breusch-Godfrey, porém, identifica-se a ocorrência de heterocedasticidade. Conforme Gujarati (2011) recomenda, estima-se o modelo de regressão com erros robustos de White para o tratamento da heterocedasticidade.

Tabela 6: Modelo de Regressão com erros Robustos de White - Modelo Original

| Coefficients | Value | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 3201.6297 | 3286.8112 | 0.9741 |
| $C$ | 1.2687 | 0.0276 | 45.9297 |
| $I$ | 1.1513 | 0.0506 | 22.7591 |
| $G$ | 0.1930 | 0.0686 | 2.8139 |
| $X$ | 1.0287 | 0.0618 | 16.6594 |
| M | -1.3141 | 0.0929 | -14.1508 |

Fonte: Elaboração própria.

Temos ainda 83 graus de liberdade e um erro residual padrão de 10010.

## 5 CONSIDERAÇÕES FINAIS

Neste trabalho, o objetivo foi estimar um modelo de regressão usando o algoritmo genético elitista, incorporando mais variáveis do que o esperado, dada a falta de conhecimento do usuário sobre as variáveis explicativas mencionadas na literatura sobre o fenômeno em estudo. O método seleciona a forma funcional do modelo com o menor valor de AIC.

Foi levado em consideração o modelo clássico de regressão linear. O qual destaca as propriedades desejáveis de um modelo de regressão, que em resumo

pode-se destacar a correta especificação do modelo e a não ocorrência da multicolinearidade, heterocedasticidade e autocorrelação.

Diversos trabalhos utilizam essa metodologia em vários ramos da ciência, inclusive na economia. Assim, a análise que foi realizada é importante para a busca de uma melhor especificação de modelos em trabalhos empíricos na literatura econômica. Foi escolhido um modelo macroeconômico simplificado que relaciona o PIB ao consumo, investimento, gastos do governo, exportações e importações.

Os resultados encontrados mostram que o modelo escolhido pelo algoritmo apresentou o menor valor do AIC (1980.991), significância estatística individual para todos os parâmetros, exceto para apenas um deles. Possui significativo poder de explicação conjunta e não apresentou erro de especificação. Porém, constatou-se no modelo, através de testes estatísticos adequados, a presença de multicolinearidade, heterocedasticidade e autocorrelação.

Em relação aos problemas identificados pelos testes realizados, caso o objetivo do modelo estimado seja apenas previsão, a ocorrência da multicolinearidade fraca não chega a ser um problema grave (GUJARATI, 2011). Além disso, foi detectado a ocorrência de heterocedasticidade e autocorrelação. Como medida de correção foi estimado um modelo de regressão com erros consistentes para heterocedasticidade e autocorrelação.

Por fim, comparando os resultados obtidos a partir das variáveis explicativas apontadas pela literatura econômica, o resultado foi a estimação de um modelo com AIC (2054.196), com menos parâmetros com significância individual, sem ocorrência de autocorrelação, porém, com erro de especificação, presença de multicolinearidade e heterocedasticidade.

Embora o presente trabalho tenha sua relevância, trata-se apenas de um trabalho inicial no qual permite que várias outras análises sejam feitas. Por exemplo, esse mesmo estudo pode ser aplicado a outras teorias econômicas na busca de um modelo mais robusto de previsão, dentre outras abordagens que ficam como sugestões de pesquisas futuras.

## REFERÊNCIAS

ACOSTA-GONZÁLEZ, E.; FERNÁNDEZ-RODRÍGUEZ, F. Model selection via genetic algorithms illustrated with cross-country growth data. Empirical economics, Springer, v. 33, n. 2, p. 313–337, 2007

ANDERSON, D. A.; BURNHAM, K. P.; ANDERSON, D. R. Model selection and inference: a practical information-theoretic approach. [S.l.]: Springer, 1998

GOLDBERG, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning. New York: Addison-Wesley, 1989.

GREENE, W. H. Econometric Analysis. Prentice Hall, 2011.

GUJARATI, Damodar N; PORTER, Dawn C. Econometria Básica. Porto Alegre: Mc Graw Hill, 2011.

HOLLAND, J. Adaptation in natural and artificial systems. Ann Arbor: The University of Michigan Press, 1975.

HURVICH, C. M.; TSAI, C.-L. Bias of the corrected aic criterion for underfitted regression and time series models. Biometrika, Oxford University Press, v. 78, n. 3, p. 499–509, 1991.

JAMES, G. et al. An introduction to statistical learning. [S.l.]: Springer, 2013. v. 112.

LACERDA, E. G. de; CARVALHO, A. C. de; LUDERMIR, T. B. Model selection via genetic algorithms for rbf networks. Journal of Intelligent & Fuzzy Systems, IOS Press, v. 13, n. 2-4, p. 111–122, 2002.

LOPES, Luiz. Martins.; VASCONCELLOS, Marco Antonio Sandoval. Manual de Macroeconomia Básico e intermediário. São Paulo Editora: Atlas, 2008.

MAIA, A. G. Econometria: conceitos e aplicações. São Paulo: Saint Paul Editora, 2017.

PATERLINI, S.; MINERVA, T. Regression model selection using genetic algorithms. In: WORLD SCIENTIFIC AND ENGINEERING ACADEMY AND SOCIETY (WSEAS). Proceedings of the 11th WSEAS international conference on rural networks and 11th WSEAS international conference on evolutionary computing and 11th WSEAS international conference on Fuzzy systems. [S.l.], 2010. p. 19–27

PEREIRA, A. et al. Convergence analysis of an elitist non-homogeneous genetic algorithm with crossover/mutation probabilities adjusted by a fuzzy controller. Chilean Journal of Statistics (ChJS), v. 9, n. 2, p. 19–32, 2018.

PEREIRA, A. G. et al. On the convergence rate of the elitist genetic algorithm based on mutation probability. Communications in Statistics - Theory and Methods, v. 49, p. 769–780, 2019.

RAO, C. R.; TOUTENBURG, H. Linear models. In: Linear models. [S.l.]: Springer, 1995. p. 3–18.

RISH, I.; GRABARNIK, G. Sparse modeling: theory, algorithms, and applications. [S.l.]: CRC press, 2014.

RUDOLPH, G. Convergence analysis of canonical genetic algorithms. IEEE Transactions on Neural Networks, 5, p. 96–101, 1994.

SUGIURA, N. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. Communications in Statistics-theory and Methods, Taylor & Francis, v. 7, n. 1, p. 13–26, 1978.