Prevendo Emissões de CO2 com dados da ONU: África do Sul

Por:
Antonio Ferreira
Eduardo Fanelli

Professor:
Paulo Maranhão

Introdução

Em um mundo cada vez mais preocupado com a preservação ambiental e o aquecimento global, se faz necessário acompanhar as taxas de emissão de gás carbônico apresentadas por cada país do planeta. Mais que isso, é importante verificar o comportamento dessas taxas ao longo do tempo e, baseado em dados de uma nação, ser capaz de prever a emissão de CO2 que ocorrerá.

A ONU coleta dados de seus países-membros acerca dos mais diversos temas, incluindo as emissões de CO2. Este trabalho visa utilizar os dados disponíveis para a África do Sul e gerar um modelo capaz de prever a emissão do gás neste país baseado em outros dados coletados.

Limpeza do Conjunto de Dados

Em posse da base de dados, foi feita uma limpeza de campos que não poderiam ser utilizados pelo modelo devido à sua baixa qualidade. Além disto, sabemos que a eficácia do modelo aumentaria com maiores quantidades de dados.

Durante a análise exploratória, observamos que a maior parte das colunas possuía 51 observações. Desta maneira, utilizamos apenas variáveis que contivessem este número para o modelo.

Após a eliminação de dados com poucas observações, foi feita uma limpeza nas colunas restantes. Excluímos então, as variáveis com ausência de muitos valores (presença de NAs). O último passo da limpeza foi a remoção de colunas que apresentassem apenas um valor (desvio padrão zero).

Ao final da etapa de limpeza, restaram 200 variáveis resposta candidatas a serem utilizadas pelo modelo.

Variáveis Explicativas: Primeira tentativa

As primeiras tentativas de definir quais variáveis seriam utilizadas no modelo foram realizadas considerando a correlação entre cada uma das 200 variáveis (X) e a variável resposta (Emissões de CO2 - Y). Mesmo considerando apenas variáveis com mais de 80% de correção com Y, poucas candidatas foram eliminadas, restando cerca de 150.

Por conseguinte, tentamos fazer uma análise de correlação entre as próprias variáveis candidatas, comparando-as duas a duas e, em caso de correlação maior que 80%, descartando a que tivesse menor correlação com Y. Este processo sugeriu a utilização de apenas uma variável explicativa devido à multicolinearidade entre elas.

Efetuado as escolhas para X, testamos o modelo, que à priori obteve sucesso. Contudo, a análise de resíduos revelou grande padronização de ruído, tornando a utilização do modelo inviável.

Tentamos então utilizar diversas combinações de transformações nas variáveis X e Y, como por exemplo raiz quadrada, log, e BoxCox. Nenhuma das combinações de transformações testadas foi capaz de reduzir a padronização de ruído observada. Decidimos, então, procurar outra maneira de selecionar as variáveis explicativas.

Análise de Componentes Principais

Levando em consideração a alta correlação entre os dados disponíveis, testamos a utilização da técnica de Análise de Componentes Principais para a construção do modelo.

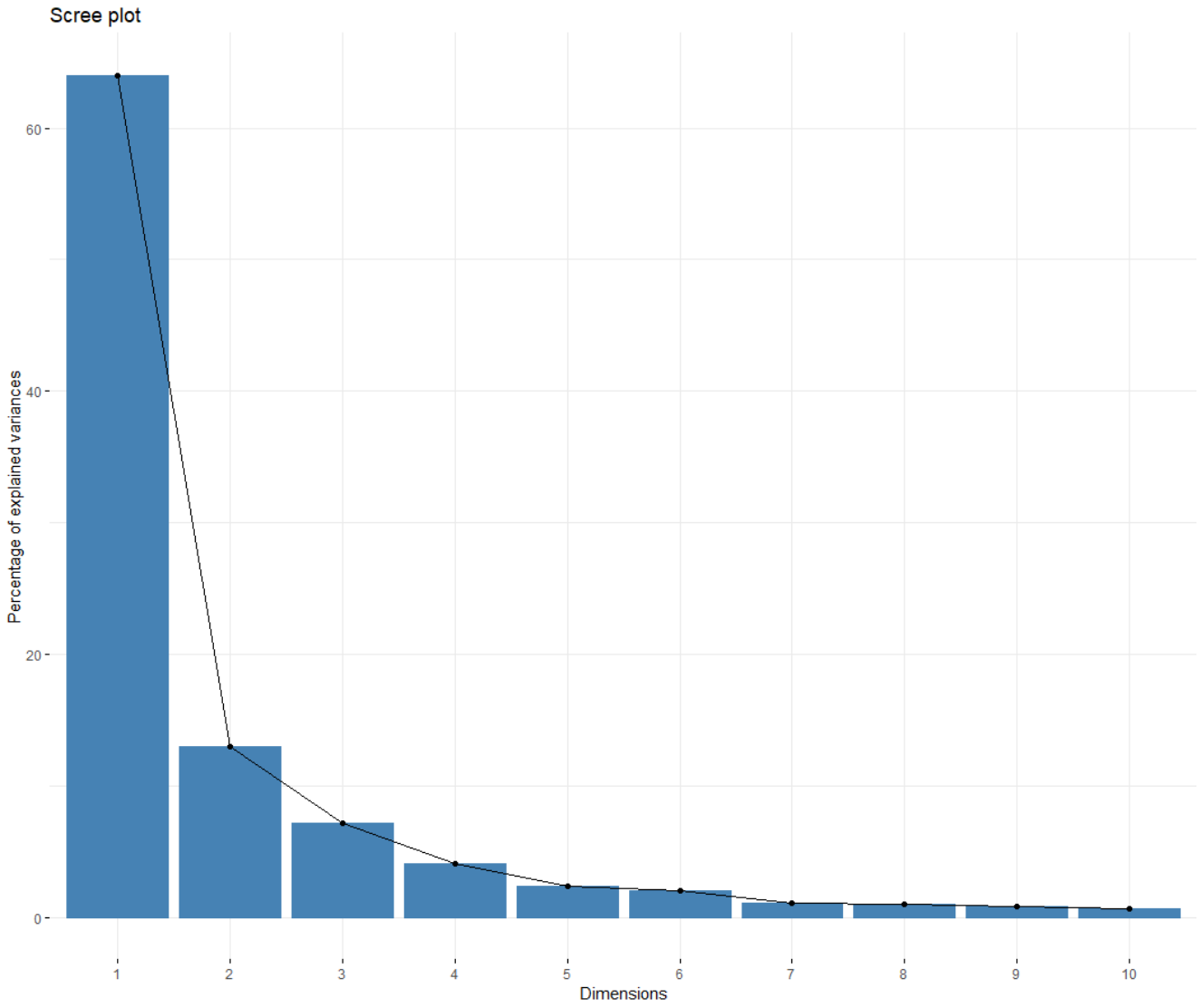
Para esta análise, utilizamos a função PRCOMP(), que divide as variáveis em grupos (PC's) de acordo com suas variâncias, formando combinações lineares.

```
anacp <-prcomp(dados_tratados, scale = TRUE)
summary(anacp)
```

O resultado do agrupamento das variáveis e suas respectivas variâncias pode ser visto abaixo:

```
Importance of components%s:
                            PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     11.3446 5.1004 3.78397 2.84785 2.17171 2.03435 1.44106 1.38698
Proportion of Variance  0.6403 0.1294 0.07124 0.04035 0.02346 0.02059 0.01033 0.00957
Cumulative Proportion   0.6403 0.7697 0.84096 0.88131 0.90477 0.92536 0.93569 0.94527
                            PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16
Standard deviation      1.31298 1.15355 1.07428 0.95296 0.92102 0.84704 0.79659 0.73982
Proportion of Variance  0.00858 0.00662 0.00574 0.00452 0.00422 0.00357 0.00316 0.00272
Cumulative Proportion   0.95384 0.96046 0.96620 0.97072 0.97494 0.97851 0.98167 0.98439
                           PC17    PC18    PC19    PC20    PC21    PC22    PC23    PC24
Standard deviation      0.65692 0.62900 0.56079 0.51939 0.49686 0.44304 0.41725 0.38972
Proportion of Variance  0.00215 0.00197 0.00156 0.00134 0.00123 0.00098 0.00087 0.00076
Cumulative Proportion   0.98654 0.98851 0.99007 0.99141 0.99264 0.99362 0.99449 0.99524
                           PC25    PC26    PC27    PC28    PC29    PC30    PC31    PC32
Standard deviation      0.37821 0.32636 0.31408 0.29533 0.27317 0.26232 0.23917 0.21624
Proportion of Variance  0.00071 0.00053 0.00049 0.00043 0.00037 0.00034 0.00028 0.00023
Cumulative Proportion   0.99595 0.99648 0.99697 0.99741 0.99778 0.99812 0.99841 0.99864
                           PC33    PC34    PC35    PC36    PC37    PC38    PC39    PC40
Standard deviation      0.20981 0.19669 0.17976 0.17145 0.15663 0.14702 0.13524 0.12568
Proportion of Variance  0.00022 0.00019 0.00016 0.00015 0.00012 0.00011 0.00009 0.00008
Cumulative Proportion   0.99886 0.99905 0.99921 0.99936 0.99948 0.99959 0.99968 0.99976
                           PC41    PC42    PC43    PC44    PC45    PC46    PC47    PC48
Standard deviation      0.10613 0.09765 0.08280 0.07340 0.06655 0.06483 0.05160 0.04283
Proportion of Variance  0.00006 0.00005 0.00003 0.00003 0.00002 0.00002 0.00001 0.00001
Cumulative Proportion   0.99981 0.99986 0.99989 0.99992 0.99994 0.99996 0.99998 0.99999
                           PC49    PC50     PC51
Standard deviation      0.03941 0.03745 1.116e-14
Proportion of Variance  0.00001 0.00001 0.000e+00
Cumulative Proportion   0.99999 1.00000 1.000e+00
```

O gráfico abaixo mostra a representatividade dos 10 primeiros componentes principais.



A alta representatividade do primeiro componente principal (>60%) evidencía a alta correlação entre muitas das variáveis que compõem a base.

Modelo Linear

Em posse dos componentes principais, foram realizados diversos testes de regressão linear. Cada teste era realizado a partir da adição de um PC ao modelo, e a constatação de se a operação contribuía ou não com o modelo.

Após várias iterações de teste, escolhemos um modelo com seis componentes principais: PC1, PC2, PC4, PC7, PC9 e PC12.

```
mod<-lm(EN_ATM_CO2E_KT~PC1+PC2+PC4+PC7+PC9+PC12, data=componentes)
```

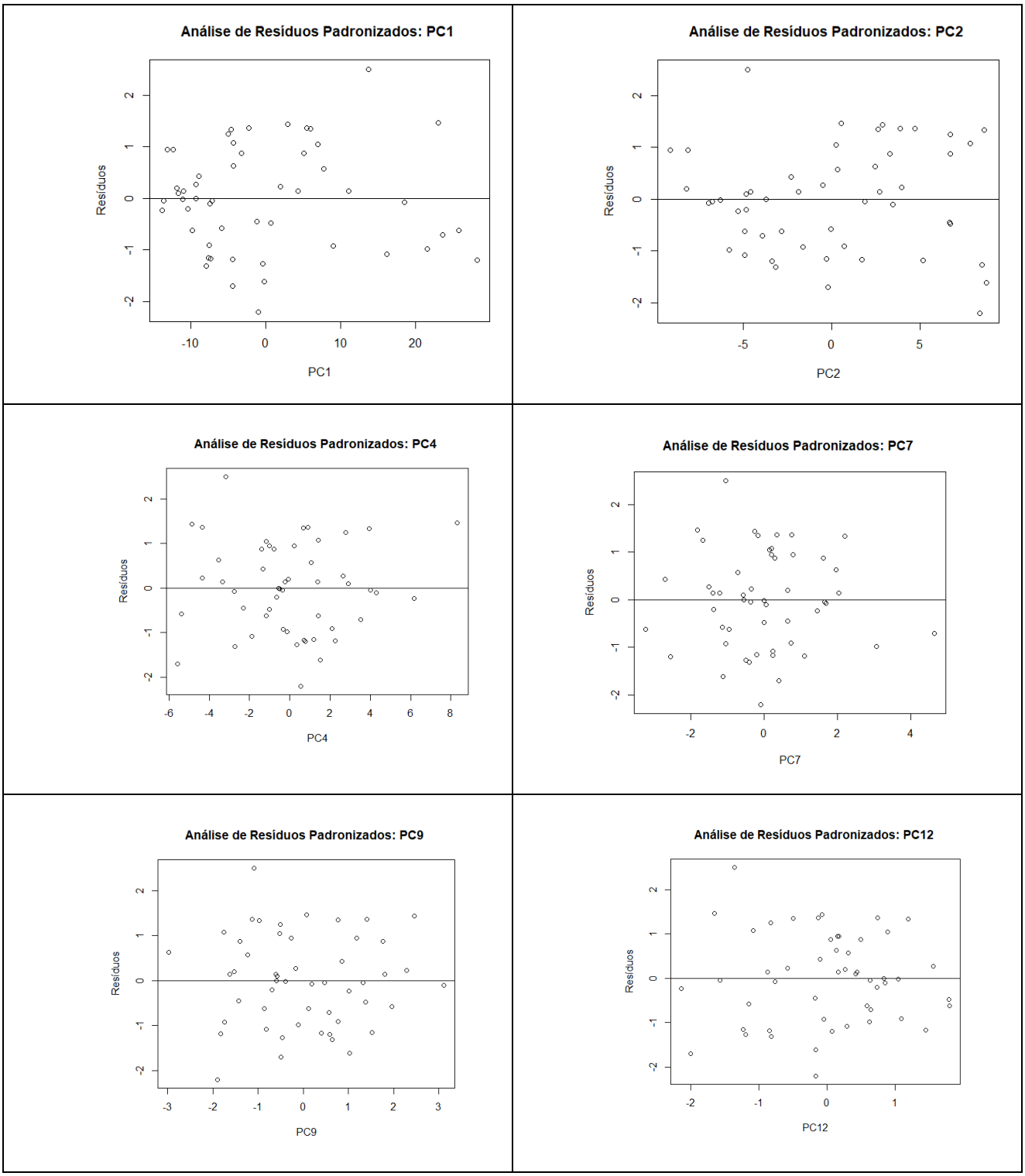
A imagem abaixo exibe as estatísticas do modelo final:

```
> summary(mod)

Call:
lm(formula = EN_ATM_CO2E_KT ~ PC1 + PC2 + PC4 + PC7 + PC9 + PC12,
    data = componentes)

Residuals:
   Min     1Q Median     3Q    Max
-29721 -10457   -221  12032  33083

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 283461.0     2018.8 140.412  < 2e-16 ***
PC1           9138.1      179.7  50.846  < 2e-16 ***
PC2           7645.1      399.7  19.125  < 2e-16 ***
PC4          -3951.9      715.9  -5.520 1.70e-06 ***
PC7           6266.5     1414.8   4.429 6.19e-05 ***
PC9          -6589.6     1552.9  -4.244 0.000112 ***
PC12         -4719.4     2139.5  -2.206 0.032666 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14420 on 44 degrees of freedom
Multiple R-squared:  0.9857,    Adjusted R-squared:  0.9837
F-statistic:   504 on 6 and 44 DF,  p-value: < 2.2e-16
```

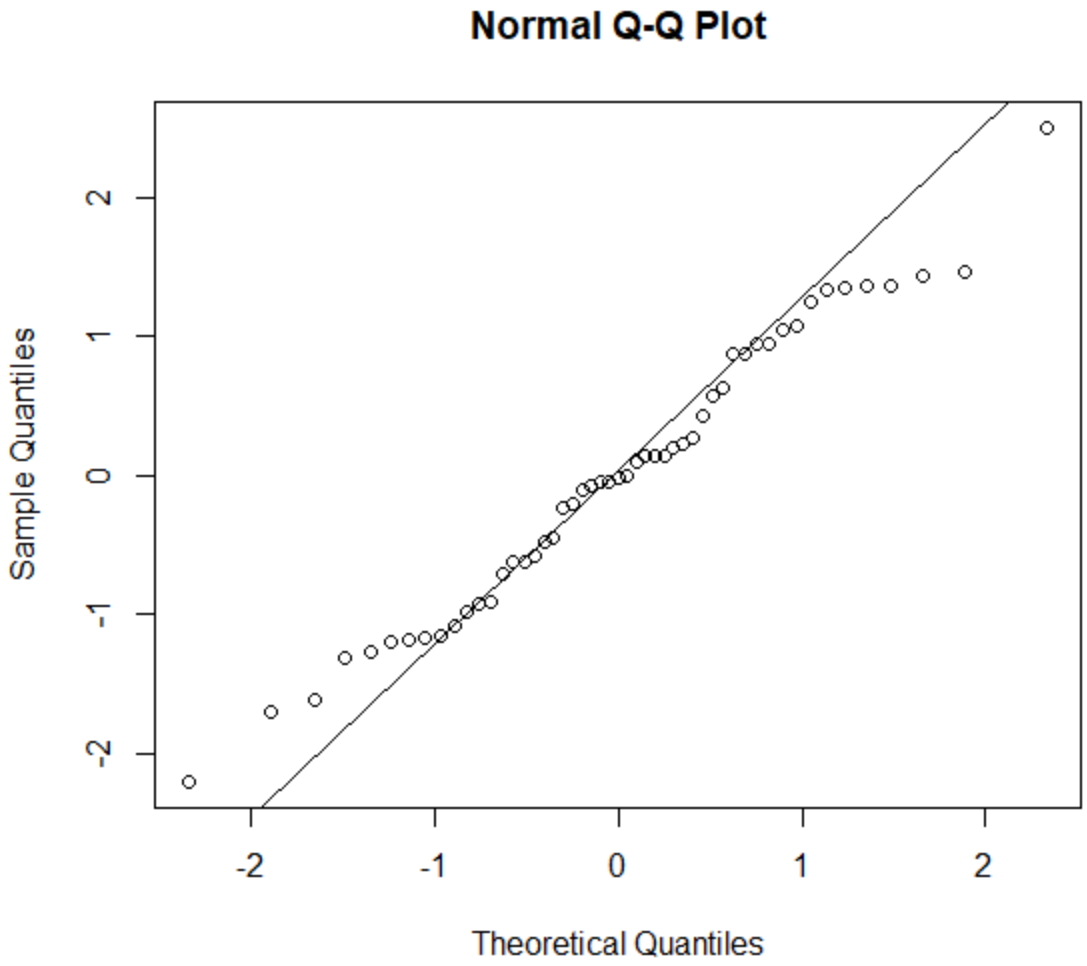
Concluímos que este modelo era promissor devido aos baixos valores de estatísticas t e alto R^2 ajustado (>98%), além do baixíssimo p-valor (< 2.2E-16).

Análise de Resíduos

Para verificar o ajuste do modelo, fizemos a análise de resíduos de cada componente principal utilizada como variável explicativa, além da variável Y:



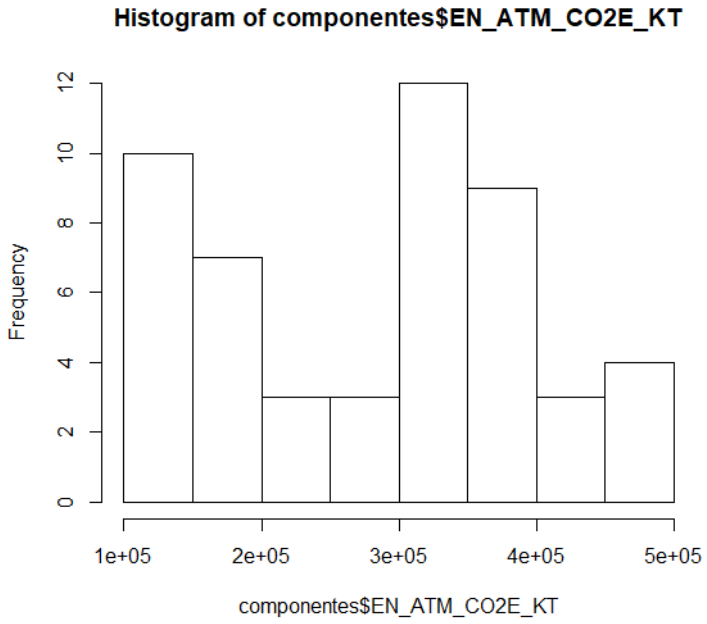
Os gráficos acima mostram a aleatoriedade dos resíduos, o que sugere um bom ajuste para o modelo. Desta forma, seguimos adiante para a construção do gráfico Q-Q, mostrado abaixo:



A análise deste plot nos diz que, apesar de existirem *outliers* o modelo se comporta razoavelmente bem para a grande maior parte dos casos.

Normalidade da Variável Y

Sabemos que uma das premissas para a adoção do método de regressão linear é que a variável dependente, Y (neste caso, emissões de CO2), deve ser normalmente distribuída. Desta maneira, apesar de conseguir um modelo aparentemente bom para os dados de treino, a usabilidade do modelo para dados adicionais depende da realização desta premissa.



O histograma acima mostra a frequência dos valores de Y. A análise visual sugere um comportamento senoidal, e não uma distribuição normal.

Para testar se podemos realmente descartar a hipótese de Y ser normal, realizamos os seguintes testes de hipótese:

```
> ad.test(componentes$EN_ATM_CO2E_KT)

        Anderson-Darling normality test

data:  componentes$EN_ATM_CO2E_KT
A = 1.1754, p-value = 0.004109

> shapiro.test(componentes$EN_ATM_CO2E_KT)

        Shapiro-Wilk normality test

data:  componentes$EN_ATM_CO2E_KT
W = 0.93484, p-value = 0.007667
```

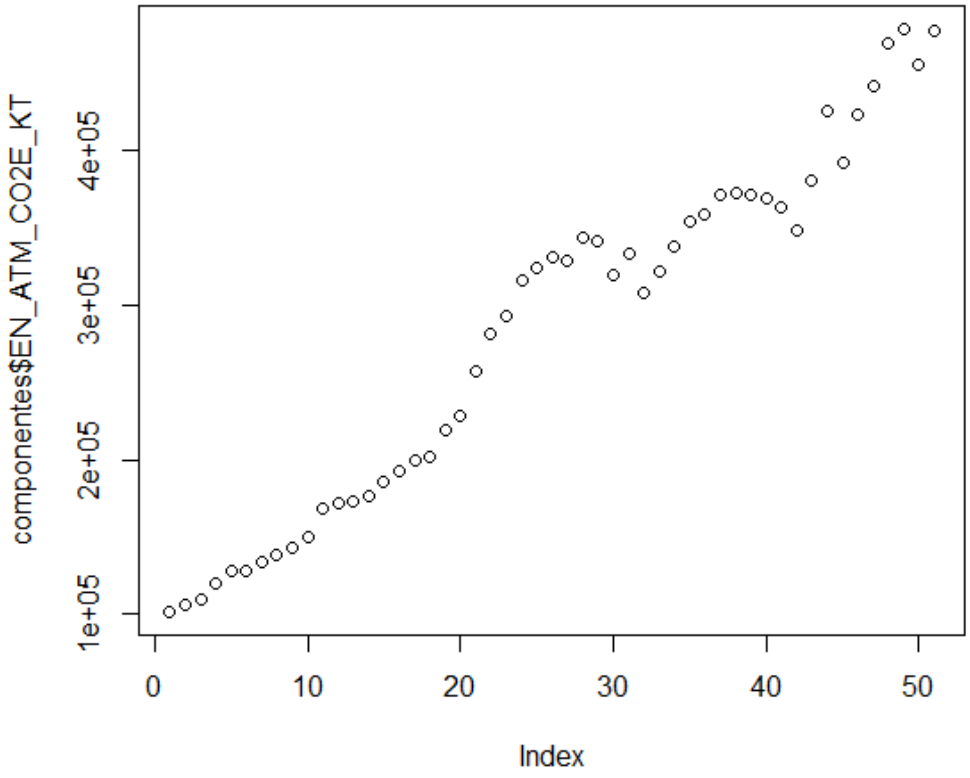
Com os baixos p-valores (<0.05), podemos rejeitar a hipótese de que a distribuição de Y seja semelhante à uma distribuição normal.

Conclusão

Neste trabalho, foram feitas diversas tentativas de construção de um modelo linear para prever a emissão de CO2 pela África do Sul. Foi constatado que este modelo seria ineficaz devido à presença de multicolinearidade entre as variáveis.

A adoção da técnica de Análise de Componentes Principais se mostrou bastante útil para resolver a multicolinearidade e possibilitou a construção de um modelo que parecia funcionar bem para os dados à mão.

Entretanto, a premissa de normalidade da variável resposta não é atendida nos dados existentes, conforme mostrado nos testes de hipótese e corroborado pelo gráfico abaixo:



Desta forma, verificamos que ao contrário de outros países, as emissões de CO2 da África do Sul seguem aumentando até o último ano da pesquisa. Como não há uma distribuição normal nesta variável, o método de regressão linear não será eficaz para prever novos valores da variável resposta fora da amostra de dados. Concluímos, portanto, que outra técnica deveria ser utilizada para atingir melhores resultados, como por exemplo Análise de Séries Temporais.