Projeto de ME731 - Análise Estatística e Modelagem do Dataset Ozone utilizando Modelo de Regressão Linear Multivariada com Análise de Componentes Principais (PCA)

Decio Miranda Filho - 236087

7 de novembro de 2023

1 Introdução

Esse presente trabalho busca modelar as variáveis do dataset Ozone do pacote Faraway do R com a aplicação da Análise de Componente Principais (PCA).Conjuntamente, será feita uma análise estatística e diagnóstica a fim de construção de um modelo de regressão linear multivariada e suas implicações.

A análise de regressão linear múltipla é uma técnica estatística amplamente utilizada para investigar a relação entre uma variável dependente e várias variáveis independentes. Ela busca modelar essa relação por meio de uma equação linear que descreve a variação da variável dependente em termos das variáveis independentes.Já o PCA (Análise de Componentes Principaiss) é frequentemente usado em modelos de regressão quando se deseja reduzir a dimensionalidade dos dados ou entender seu efeito sobre o modelo. As componentes principais são usadas como variáveis independentes em vez das variáveis iniciais. Isso pode ser útil quando há multicolinearidade entre as variáveis originais ou quando se deseja reduzir o número de preditores no modelo.

2 Descrição Matemática da Regressão Multivariada e do PCA

2.1 Regressão Multivariada

A forma geral de um modelo de regressão linear múltipla é dada pela seguinte equação:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{np-1} + \varepsilon_i$$

Onde:

- Y é a variável dependente que desejamos prever;
- $X_1, X_2, \ldots, X_p$ são as variáveis independentes que supomos estarem relacionadas a Y;
- $\beta_0, \beta_1, \ldots, \beta_p$ são os coeficientes de regressão que medem a contribuição de cada variável independente no valor de Y;
- ε é o termo de erro que representa a parte da variação de Y que não pode ser explicada pelas variáveis independentes.

O objetivo da análise de regressão linear múltipla é estimar os coeficientes de regressão $\beta_0, \beta_1, \ldots, \beta_p$ com base nos dados disponíveis e avaliar a qualidade do ajuste do modelo aos dados considerando a aplicação do PCA. Isso também envolverá a significância estatística dos coeficientes, diagnosticar a presença de violações de pressupostos.

2.2 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é uma técnica estatística usada para reduzir a dimensionalidade de um conjunto de dados enquanto preserva a maior parte da variabilidade. É amplamente aplicada em várias áreas, como análise de dados,Estatística e aprendizado de máquina.Suponha que tenhamos um conjunto de dados com n observações e p variáveis. A matriz de dados X é de dimensão $n \times p$, onde cada linha representa uma observação e cada coluna representa uma variável.

Primeiro, padroniza-se os dados subtraindo a média de cada variável e dividindo pelo desvio padrão, obtendo a matriz padronizada

Z é a matriz de dados padronizados, X é a matriz de dados original, μ é o vetor das médias das variáveis, e σ é o vetor dos desvios padrão das variáveis.

Em seguida, calcula-se a matriz de covariância das variáveis padronizadas Z. A matriz de covariância C é de dimensão $p \times p$ e é dada por:

$$C = \frac{1}{n-1} Z^T Z$$

Onde: C é a matriz de covariância, n é o número de observações, e Z^T é a matriz transposta de Z.

Em seguida, calcula-se os autovetores e autovalores da matriz de covariância C. Os autovetores representam as direções principais (componentes principais) dos dados, e os autovalores representam a quantidade de variância explicada por cada componente principal. Os autovetores são denotados como $v_1, v_2, \ldots, v_p$, e os autovalores como $\lambda_1, \lambda_2, \ldots, \lambda_p$.

Ordenamos os autovalores em ordem decrescente e selecionamos as k maiores componentes principais que explicam a maior parte da variabilidade dos dados. Tipicamente, escolhemos um valor k com base na quantidade de variância que desejamos manter.

Para obter as novas coordenadas das observações nos novos eixos (componentes principais), multiplicamos a matriz de dados padronizados Z pelos autovetores selecionados:

$$T = ZV$$

Onde: T é a matriz das observações nas novas coordenadas, Z é a matriz de dados padronizados, V é a matriz dos autovetores selecionados.

3 Análise Descritiva e Exploratória

3.1 Informações Sobre o Pacote e Variáveis

Utilizando um conjunto de dados contendo medições meteorológicas diárias coletadas pela Base da Força Aérea de Sandburg na região da Bacia de Los Angeles durante o ano de 1976. O objetivo dessa coleta foi investigar a associação entre a concentração de ozônio (O3) na atmosfera e outras variáveis meteorológicas.

Os dados utilizados foram obtidos a partir do banco de dados "ozone", disponível no pacote "faraway"do software R. Esses dados contêm informações sobre as variáveis meteorológicas e as respectivas concentrações de ozônio registradas.

Vale notar que o dataset possui **n=330** e **p=10** sendo n o número de observações e p a quantidade de variáveis presentes no dataset.

As variáveis no modelo de regressão linear múltipla são classificadas da seguinte forma:

- A variável **O3** representa a concentração de ozônio (pmm) e é considerada contínua.
- A variável **vh** representa a altura de 500 m de Vandenburg (m) e é considerada contínua.
- A variável **wind** representa a velocidade do vento (mph) e é considerada contínua.
- A variável **humidity** representa a umidade em porcentagem (%) e é considerada contínua.
- A variável **temp** representa a temperatura em graus Celsius (°C) e é considerada contínua.
- A variável **ibh** representa a altura em pés (ft.) e é considerada contínua.
- A variável **dpg** representa o gradiente de pressão Daggett (mmHg) e é considerada contínua.

- A variável **ibt** representa o índice de invesão base da temperature LAX.
- A variável **vis** representa a visibilidade em milhas e é considerada contínua.
- A variável **doy** representa o dia do ano (1976) e é considerada discreta.

Essa classificação é importante para a análise e interpretação do modelo de regressão linear múltipla utilizado no estudo.

3.2 Descrição e Visualização

Variável	Min	Mediana	Média	Max
O3	1.00	10.00	11.78	38.00
vh	5320	5760	5750	5950
wind	0.000	5.000	4.848	11.000
humidity	19.00	64.00	58.13	93.00
temp	25.00	62.00	61.75	93.00
ibh	111.0	2112.5	2572.9	5000.0
dpg	-69.00	24.00	17.37	107.00
ibt	-25.0	167.5	161.2	332.0
vis	0.0	120.0	124.5	350.0
doy	33.0	205.5	209.4	390.0

A tabela acima nos mostra as principais estatísticas descritivas das variáveis.O3 (Ozônio): A concentração média de ozônio é de aproximadamente 11.78 ppm, com valores variando de 1.00 ppm a 38.00 ppm. A mediana é de 10.00 ppm, indicando uma distribuição assimétrica à direita. Essa variável pode ser considerada como a variável é a resposta do nosso modeloA velocidade média do vento é de aproximadamente 5750, com valores variando de 5320 a 5950. A direção do vento possui uma média de 4.848 e varia de 0.000 a 11.000. Essa variável indica a direção do vento e pode ser útil para entender o padrão de dispersão do ozônio na região.A umidade média é de aproximadamente 58.13, com valores variando de 19.00 a 93.00. A umidade pode influenciar na concentração e dispersão do ozônio na atmosfera.A temperatura média é de aproximadamente 61.75, com valores variando de 25.00 a 93.00. A temperatura pode ter uma relação com a formação e degradação do ozônio na atmosfera.O valor médio da inversão da base da nuvem é de aproximadamente 2572.9, com valores variando de 111.0 a 5000.0.

Tabela 1: Estatísticas Descritivas da Variável O3

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	5.00	10.00	11.78	17.00	38.00

A partir das estatísticas descritivas, observa-se que a concentração média de ozônio foi de aproximadamente 11.78 ppm A mediana da variável O3 foi de 10.00 ppm, indicando uma distribuição ligeiramente assimétrica à direita.Os quartis 1 e 3 (*1st Qu.* e *3rd Qu.*) foram calculados em 5.00 ppm e 17.00 ppm, respectivamente. Esses valores representam os limites inferiores e superiores dos 25% e 75% dos dados, indicando a dispersão dos valores em torno da mediana.

Segundo a figura acima conseguimos perceber as distribuições por frequência das variáveis, exceto a resposta. Conjuntamente com o plot da matriz correlações entre todas as variáveis, uma a uma, haverá uma seleção de variáveis para a construção inicial do modelo na próxima etapa. Percebe-se que a variável vh e ibt possuem alta correlação, o que pode ser um indício de um problema de multicolinearidade. Isso se repete também para o conjunto temp,ibt e vh,temp.

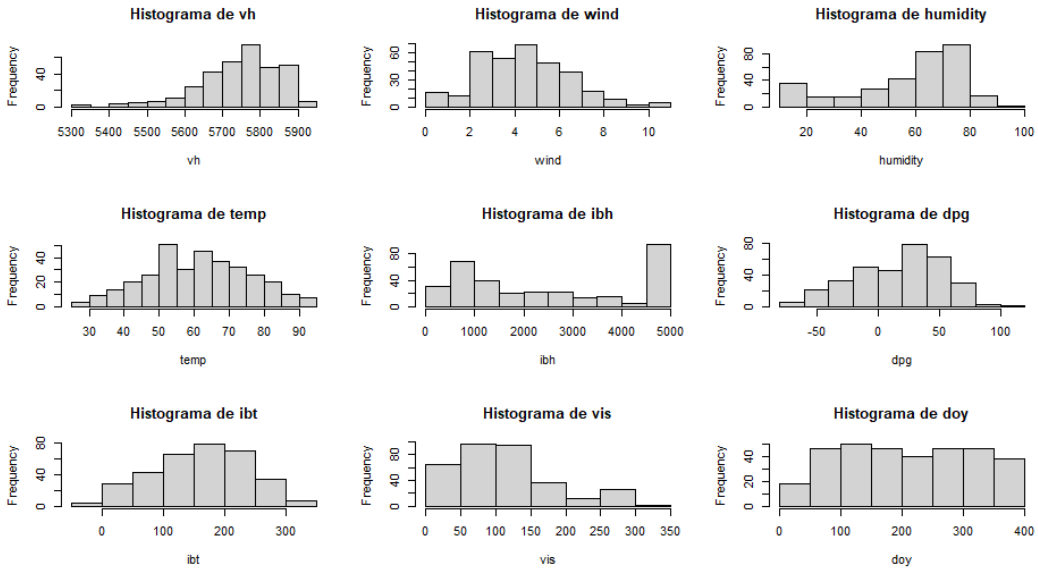

Figura 1: Distribuição das Frequências das Variáveis Regressoras.

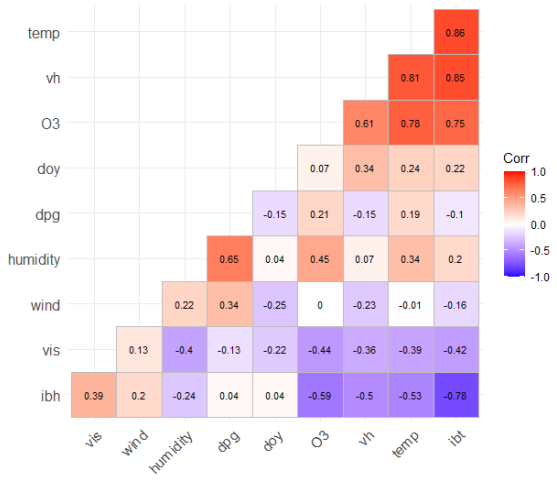

Figura 2: Matriz de Correlação entre as Variáveis

4 Construção do Modelo

Para esse trabalho será utilizada a significância estatística $\alpha = 5\%$

4.1 Padronização da Escala

Para o PCA, escalonar variáveis é crucial para equilibrar sua influência na análise, especialmente quando elas têm escalas diferentes. O escalonamento,no R, assegura que todas as variáveis contribuam igualmente, evitando que variáveis de maior escala dominem a análise. Isso simplifica a interpretação e ajuda a identificar padrões precisos nos dados.

4.2 Regressão Múltipla e Aplicação do PCA

Como medida de comparação, num primeiro momento realizar-se-á a regressão linear múltipla sem a aplicação do PCA para medida de comparação e análises.

Após a regressão,a principal diferença entre as duas regressões está nos coeficientes. Com a aplicação

do PCA, os coeficientes não representam diretamente as variáveis originais, mas sim as relações entre os componentes principais e a variável dependente. Isso é uma consequência do PCA, que transforma as variáveis originais em novas variáveis não correlacionadas.

No entanto, observe que as estatísticas de ajuste, como o R-quadrado 0.6967, e a estatística F: 92.18, são equivalentes em ambas as regressões. Isso indica que ambas as regressões, explicam a mesma quantidade de variação na variável dependente e não há diferença significativa no ajuste do modelo.

4.3 Interpretação do PCA e Redução de Dimensionalidade

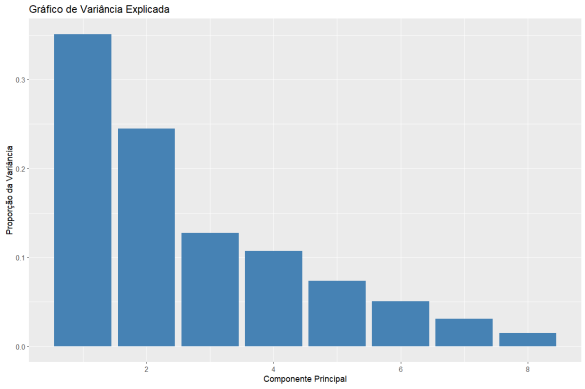

Figura 3: Gráfico do Percentual das Variâncias Explicadas das Componentes Principais

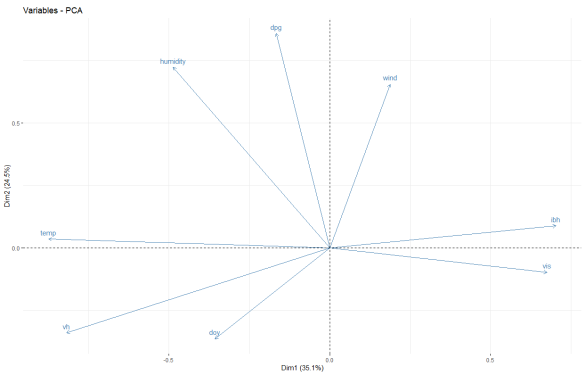

Figura 4: Biplot das Variáveis Principais

O gráfico de variância explicada em uma análise de PCA é uma representação visual que ajuda a entender quanto de variabilidade nos dados é explicado por cada componente principal. Cada barra no gráfico representa um componente principal, e a altura da barra indica a proporção da variância total nos dados que aquele componente captura. A interpretação dos valores é essencial para decidir quantos componentes reter e como esses componentes contribuem para a compreensão dos dados. Ao lado da figura de variância explicada há o plot que destaca variáveis originais no espaço de componentes principais. A posição relativa dos pontos mostra como as variáveis se relacionam. Pontos próximos indicam correlações, e aqueles que se estendem na direção de uma componente têm alta contribuição. Ajuda a identificar variáveis mais associadas às componentes principais.

Mediante isso, observa-se que As **Componentes Principais (PC1,PC2,PC3,PC4):** São responsáveis por capturar cerca de 35,11%,24,45%,12,74% e 10,72% da variabilidade total nos dados, respectivamente. PC1 é a componente mais importante e está associada aos padrões dominantes nos dados ao passo que é a segunda componente mais relevante, representando variação adicional que é ortogonal a PC1. PC3 e PC4 já apresentam uma variabilidade menor ainda que significante.

Estas são as componentes mais importantes e que explicam a maior parte da variação nos dados. Em muitos casos, reter essas PCAs será suficiente para preservar as informações essenciais enquanto reduz a dimensionalidade dos dados. A interpretação das PCAs subsequentes pode ser realizada, mas elas geralmente explicam proporções menores da variação e são menos críticas para a compreensão dos padrões nos seus dados.

4.4 Modelagem a Partir da Redução do PCA

Após a análise da variância explicada por componentes principais e testes diretos no R decidiu-se utilizar as 5 maiores componentes para realização da regressão multivariada.

A regressão tem a seguinte fórmula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot PC_1 + \hat{\beta}_2 \cdot PC_2 + \ldots + \hat{\beta}_5 \cdot PC_5 + \hat{\epsilon}$$

Os coeficientes estimados associados para o modelo são: - Intercepto: 11.7758 (p-valor muito baixo, indicando significância estatística) - PC1: -3.6906 (p-valor muito baixo, indicando significância

estatística) - PC2: 0.8744 (p-valor muito baixo, indicando também significância estatística) - PC3: -1.5703 (p-valor muito baixo, indicando também significância estatística) - PC4: 0.8171 (p-valor baixo, indicando também significância estatística) - PC5: 0.4087 (p-valor alto, não é estatisticamente significativo)

O R^2 é cerca de 66%, o que significa que aproximadamente 66% da variabilidade na variável de resposta Y é explicada pelos componentes principais. O valor F é 131 e o p-valor aproximadamente zero, o que indica que o modelo de regressão com os componentes principais é estatisticamente significativo.

A tabela ANOVA forne ce informações sobre a variação explicada por cada componente principal e a variação residual. Cada componente principal (PC1, PC2, PC3, PC4, PC5) tem um valor F associado com um p-valor muito baixo, o que indica que cada componente principal é significativo na explicação da variação na variável de resposta Y. A maior parte da variação é explicada por PC1, seguida por PC2, PC3, e assim por diante.

O modelo de regressão com PCA parece ser estatisticamente significativo e explicar uma parte substancial da variabilidade na variável de resposta Y. Os componentes principais PC1, PC2 e PC3 têm as maiores contribuições na explicação da variação em Y, enquanto PC4 e PC5 têm contribuições menores. Essa seleção de componentes principais e seus coeficientes fornecem informaçõs sobre como as variáveis originais estão relacionadas à variável de resposta após a redução de dimensionalidade.

4.5 Diagnósticos dos Resíduos

Com a análise dos reísudos os testes sugerem que os resíduos do modelo com PCA não seguem estritamente uma distribuição normal, com um p-valor de 0.04394 no teste de Shapiro-Wilk. Além disso, há evidências de heterocedasticidade, com um p-valor muito próximo de zero (1.543e-06) no teste Breusch-Pagan.

4.6 Transformação por Box-Cox

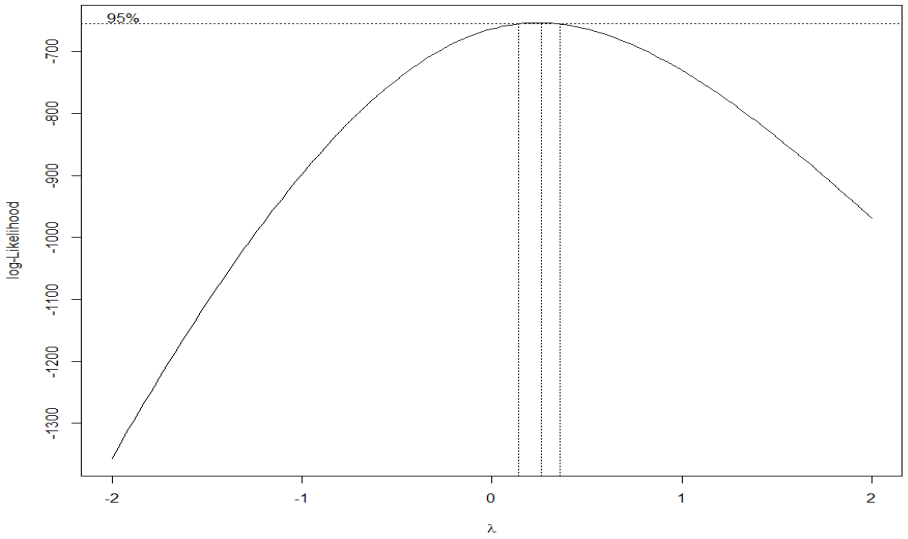

Figura 5: Distribuição das Frequências das Variáveis Regressoras.

A transformação de Box-Cox é uma ferramenta valiosa na análise de regressão quando os diagnósticos dos resíduos indicam problemas de normalidade e heterocedasticidade. Testando valores próximos de lambda que maximizem a verossimilhança identifica-se que quando o lambda λ da transformação se aproxima de 1/2 isso ocorre.Assim, aplica-se uma transformação de raiz quadrada sobre a variável dependente. Essa transformação é útil para estabilizar a variância dos resíduos e, ao mesmo tempo, torná-los mais próximos de uma distribuição normal.

A transformação de raiz quadrada $\lambda = 1/2$ é especialmente útil quando os resíduos exibem heterocedasticidade, pois ela ajuda a reduzir a variação da variância ao longo do espaço amostral. Além disso, a transformação pode tornar os resíduos mais simétricos e próximos de uma distribuição normal, melhorando a validade dos pressupostos da regressão linear.

Portanto, estará abordando efetivamente problemas de heterocedasticidade e normalidade nos resíduos, tornando seus resultados de regressão mais confiáveis e interpretáveis.

4.7 Remodelando a partir da transformação

Logo, o modelo finalo será dado por:

$$\sqrt{\hat{y}} = 3.23295 - 0.54288 \cdot PC_1 + 0.13356 \cdot PC_2 - 0.25356 \cdot PC_3 + 0.10181 \cdot PC_4 + 0.07726 \cdot PC_5 + \hat{\epsilon} \quad (1)$$

O modelo de regressão foi ajustado usando a transformação de raiz quadrada de O3 como a variável de resposta. Os coeficientes estimados para o modelo são os seguintes:

- Intercepto: 3.23295 (p-valor muito baixo, altamente significativo) - d_box$PC1: -0.54288 (p-valor muito baixo, altamente significativo) - d_box$PC2: 0.13356 (p-valor muito baixo, altamente significativo) - d_box$PC3: -0.25356 (p-valor muito baixo, altamente significativo) - d_box$PC4: 0.10181 (p-valor baixo, significativo) - d_box$PC5: 0.07726 (p-valor alto, não é estatisticamente significativo) Resíduos:

Os resíduos do modelo têm uma média próxima de zero e uma variação bastante baixa, indicando que a transformação de Box-Cox foi eficaz na estabilização da variância dos resíduos. A distribuição dos resíduos parece se aproximar de uma distribuição normal, com valores mínimos e máximos próximos a zero.O R^2 ajustado é de 0.704, o que significa que aproximadamente 70,4% da variabilidade na raiz quadrada de O3 é explicada pelos componentes principais e outras variáveis independentes.O valor de F é 157.5 com um p-valor próximo de zero, indicando que o modelo de regressão é altamente significativo.

A tabela ANOVA fornece informações sobre a variação explicada por cada componente principal e a variação residual. - Cada componente principal tem um valor F associado com um p-valor muito baixo, o que indica que cada componente principal é significativo na explicação da variação na raiz quadrada de O3.

4.8 Diagnósticos Pós Transformação

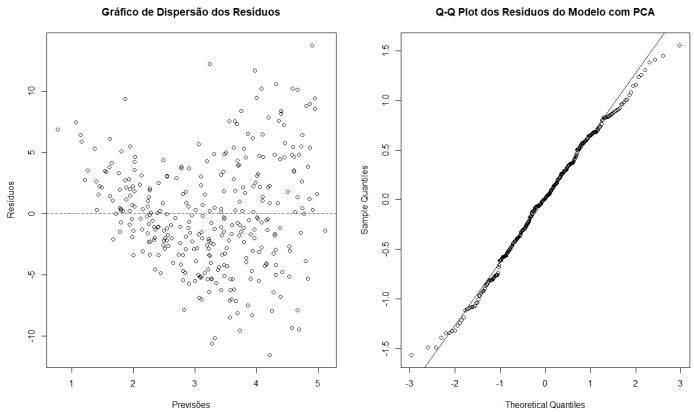

Figura 6: Análise de Resíduos

Após a transformação de Box-Cox, os resultados dos testes de normalidade e homocedasticidade dos resíduos são os seguintes:

O teste de normalidade de Shapiro-Wilk para os resíduos apresenta um valor de W = 0.99313 e um p-valor de 0.1363. O resultado indica que os resíduos não diferem significativamente de uma distribuição normal, sugerindo uma melhoria na normalidade após a transformação de Box-Cox.

O teste de Breusch-Pagan para heterocedasticidade mostra um valor de BP = 5.2264 com um p-valor de 0.3889. O p-valor relativamente alto indica que não há evidências significativas de heterocedasticidade nos resíduos.

Esses resultados sugerem que a transformação de Box-Cox foi eficaz na correção das não normalidades e na estabilização da variância dos resíduos. O modelo de regressão parece cumprir melhor os pressupostos da regressão linear após a transformação.

5 Conclusão

O PCA proporcionou benefícios significativos na análise dos dados de poluição do ar nos EUA, reduzindo a complexidade para cinco dimensõe principais e explicando grande parte da variabilidade original, facilitando a compreensão dos padrões. Já a aplicação da regressão linear múltipla em conjunto com a transformação de Box-Cox possibilitou uma adequação da análise dos resíduos ao passo que se mostrou altamente siginificativo. Em resumo, a aplicação do PCA na regressão linear múltipla é uma ferramenta poderosa para reduzir a dimensionalidade dos dados. Contudo, é importante utilizá-la com cautela, além de realizar análises adicionais para garantir que o modelo resultante seja confiável e útil para a finalidade específica. A interpretabilidade das componentes principais e o tratamento de outliers são áreas que podem ser aprimoradas para obter resultados mais robustos e significativos. A aplicação do PCA na modelagem de regressão linear apresenta vantagens e desvantagens, e há oportunidades para melhorias e considerações adicionais

6 Críticas

Vantagens:

O PCA ajuda a reduzir a dimensionalidade dos dados, mantendo as informações essenciais. Isso simplifica a modelagem, tornando-a mais eficient. O PCA pode lidar com multicolinearidade, um problema comum em regressão linear quando as variáveis independentes estão altamente correlacionadas. Além disso, as componentes principais são combinações lineares não correlacionadas das variáveis originais, tornando a interpretação dos coeficientes mais simples. Dito isto, ele ajuda a identificar e remover variáveis com baixa contribuição na explicação da variabilidade, resultando em modelos mais enxutos.

Desvantagens:

Embora o PCA melhore a interpretabilidade dos coeficientes, a interpretação das componentes principais em si pode ser desafiadora em contextos mais práticos além de que redução de dimensionalidade pode resultar na perda de informações valiosas contidas nas variáveis originais. Vale notar também que o PCA é sensível a outlier e escalas diferentes, e a presença dessesnos dados pode afetar negativamente as análises sendo necessária uma adequação inicial. Além disso, perda de detalhes sutis ao reduzir a dimensionalidade é uma desvantagem pela perda de informação mesmo que reduza suas informações.

O que pode ser melhorado:

Após a aplicação do PCA, é importante realizar uma validação do modelo para garantir que ele atenda aos pressupostos da regressão linear. A interpretabilidade das componentes principais pode ser um desafio, assim se sugere uma melhoria nesse aspecto envolve o uso de técnicas de interpretação. O tratamento adequado de outliers é crucial. Técnicas robustas de PCA podem ser aplicadas para reduzir a influência de outliers.

Para melhorar, recomenda-se padronizar ou escalonar as variáveis , explorar mais dimensões adicionais, testar o PCA com outras técnicas(como correlações canônicas, visto recentemente em aula) e considerar a especificidade de cada dataset.

7 Referências e Bibliografia

Notas de aula

Everitt, B. and Dunn, G.. Applied Multivariate Data Analysis. Segunda Edição, Wiley Sons, nova Iorque, 2001

Koch, I.. Analysis of Multivariate and High Dimensional Data: Theory and Pratice. Cambridge University Press, Nova Iorque, 2014

Neter, J.; Wasserman, W. & Kutner, M. H.; Applied Linear Statitiscal Models. Quinta Edição. Irwin, Boston, 200

Mardia, K.V., Kent, J.T. and Bibby, J.M.. Multivariate Analysis. Sétima Reimpressão. Academic Press, Londres, 2000

Mood, A. M.; Graybill, F. A. Boes, D. C.; Introduction to the Theory of Statistics. Terceira Edição. McGraw-Hill, Nova Iorque, 1974.

Wickham, H.; Navarro, D. Pedersen, T. L.; ggplot2 : elegant graphics for data analysis. Terceira Edição. Springer, 2020. Disponível em: https://ggplot2-book.org. Acesso em: 7 nov. 2023.

Faraday, J. J.; Linear Models with R. Florida: Chapman Hall/CRC, v. 63, 2005

Dataset disponível no R ou em: https://search.r-project.org/CRAN/refmans/fdm2id/html/ozone.html

RDrr.io. Ozone1 dataset. Disponível em: ¡https://rdrr.io/cran/earth/man/ozone1.html¿. Acesso em: 7 nov. 2023.