CATEGORIA (1)

Aplicação da Análise de Componentes Principais ao Modelo de Divisão Modal

INTRODUÇÃO

O objetivo deste trabalho é apresentar um método para mitigar a multicolinearidade em modelos preditivos de divisão modal.

A análise dos comportamentos que influenciam as escolhas modais entre pares de origens e destinos é uma etapa crucial no planejamento de transportes. Essa etapa visa prever o comportamento da demanda com base em suas características socioeconômicas, demográficas e de acessibilidade. Segundo Willumsen (2001), a escolha modal é a questão mais importante no planejamento de transportes.

As políticas públicas baseadas em estimativas sobre a divisão modal possibilitam fomentar planos de linhas e de redes de transporte coletivos, reduzindo acidentes, emissões de gases poluentes e congestionamentos, além de promover um uso mais eficiente do espaço urbano.

Esses benefícios contribuem diretamente para o alcance dos recentes objetivos traçados pela ONU relacionados ao desenvolvimento sustentável, especialmente no que se refere a medidas contra as mudanças climáticas.

Para realizar tais estimativas o Metrô de São Paulo desenvolve modelos de previsão de escolhas modais que visam reproduzir os comportamentos da demanda em seus deslocamentos diários.

Esses modelos são aprimorados a cada edição da Pesquisa Origem Destino realizada pelo Metrô de São Paulo, e procura acompanhar a complexa dinâmica de comportamento da demanda ao longo do tempo.

A análise dos padrões de viagens intrinsecamente relacionados à divisão modal, e o desenvolvimento de novos modelos capazes de reproduzir a realidade do comportamento da demanda na RMSP após o Covid19, são os próximos desafios a serem equacionados a partir dos dados coletados na Pesquisa OD 23, em andamento pelo Metrô de São Paulo.

Os modelos de divisão modal baseados em regressão logística múltipla são amplamente utilizados por agências de transporte em todo o mundo e pelo Metrô de São Paulo. A validação desses modelos exige o atendimento de um conjunto de pressupostos estatísticos. Um desses pressupostos é a ausência de multicolinearidade entre as variáveis preditoras, ou seja, tais variáveis só podem ser aceitas se não representarem combinações lineares de outras variáveis preditoras, ou, em outras palavras, não devem ser fortemente correlacionadas entre si.

No caso de existência de multicolinearidade, a variância dos coeficientes do modelo pode se tornar muito elevada, tornando os coeficientes do modelo pouco

confiáveis ou, em casos extremos, impossibilitando a obtenção do modelo (Kleinbaum e Klein, 2010).

Para ajudar a mitigar esse problema pode ser utilizada uma técnica estatística de análise multivariada conhecida como Análise de Componentes Principais (ACP), que possibilita obter novas variáveis (dimensões) resultantes de combinações lineares não relacionadas das variáveis originais.

No presente artigo, é investigada a mitigação da multicolinearidade das variáveis preditoras de um modelo de logit binomial de divisão modal, elaborado a partir de dimensões resultantes da Análise de Componentes Principais de variáveis extraídas da Pesquisa OD 2017.

DIAGNÓSTICO

A partir dos dados da Pesquisa Origem Destino 2017 realizada pela Companhia do Metropolitano, com abrangência na Região Metropolitana de São Paulo (RMSP), foram levantadas informações de 116 mil domicílios e obtidos dados válidos para 32 mil domicílios selecionados de forma aleatória nos 39 municípios da RMSP distribuídos em 517 zonas de pesquisa.

As escolhas modais consideradas para a elaboração do modelo do presente artigo foram: transporte coletivo e o transporte individual. Para o transporte coletivo foram consideradas as viagens realizadas pelos modos metrô, trens, ônibus e microônibus do município de São Paulo, ônibus e microônibus dos demais municípios da RMSP e ônibus e microônibus metropolitanos. Para o

transporte individual foram consideradas as viagens realizadas com automóvel ou motocicleta.

Neste estudo foram consideradas as seguintes variáveis preditoras: (1) Renda Familiar na zona de origem (RFAM_O), (2) Taxa de Motorização Individual (TMI), (3) Taxa de População da classe A na origem (TX_POPA_O), (4) Renda per capita na origem (RPC_O), (5) Taxa população AB na origem (TX_POPAB_O), (6) Taxa de população C na origem (TX_POPC_O), (7) Taxa de população CDE na origem (TX_POPCDE_O), (8) população com idade de 7 a 17 anos no destino (P7a17_D), (9) população com idade 15 a 17 anos no destino (P15a17_D), (10) População no destino (POP_D), (11) População classes CDE no destino (POP_CDE_D), (12) Matrículas 1º. Grau no destino (M1GRAU_D), (13) Matrículas em escolas públicas no destino (MPUB_D), (14) População classes DE no destino (POP_DE_D), (15) População com idade 7 a 17 anos na origem (POP7a17_O), (16) Matrículas em creches na origem (MCRE_O), (17) Matrículas em 1º. Grau na origem (M1GRAU_O), (18) População das classes CDE na origem (POP_CDE_O), (19) Matrículas 2º. Grau na origem (M2GRAU_O), (20) População das classes DE na origem (POP_DE_O), (21) Taxa matrículas em curso superior no destino (TX_MSUP_D), (22) Taxa de matrículas no ensino particular no destino (TX_MPAR_D), (23) Taxa de matrículas no ensino público no destino (TX_MPUB_D), (24) Taxa de matrículas no 1º. Grau no destino (TX_M1GRAU_D), (25) Taxa de matrículas no ensino superior na origem (TX_MSUP_O), (26) Taxa de matrículas no ensino público na origem (TX_MPUB_O), (27) Taxa de matrículas no ensino particular na origem

(TX_MPAR_O), (28) Taxa de população com idade entre 15 e 17 anos na origem (TX_P15a17_O) e (29) Taxa de população com idade entre 7 e 17 anos (TX_P7a17_O).

O modelo de divisão modal com técnica de regressão logística

O tipo de técnica utilizada par amodelar a divisão modal foi a técnica de regressão logística binomial. Essa técnica apresenta algumas vantagens sobre outras técnicas, tais como, possibilita obter a probabilidade binomial de escolha entre transporte coletivo e individual, possibilita a utilização de variáveis preditoras quantitativas ou qualitativas (restritas tipicamente a categorias de valores que são mutuamente exclusivos, tais como, 2, 3 ou 4, tipicamente); não exige relação linear entre a variável dependente e as covariáveis; não exige que as variáveis independentes apresentem distribuição normal; caracteriza-se como menos sensível a valores outliers.

Em (FÁVERO e col., 2009) encontramos as seguintes premissas para a regressão logística:

- Relação linear entre o logit da variável resposta e as variáveis explicativas.
- Os resíduos devem apresentar valor esperado igual a zero.
- Ausência de heterocedasticidade.
- Ausência de multicolinearidade.

O interesse particular do presente artigo é a abordagem da questão da multicolinearidade na construção de modelos de regressão logística para a predição da divisão modal.

Seguindo a notação de (FÁVERO e col., 2009), o modelo de regressão logit apresenta a seguinte função logística $\boldsymbol{f(Z) = \frac{1}{1+e^{-Z}}}$. Para qualquer valor $\boldsymbol{Z} \in (-\infty, \infty)$ a função logística assume valores entre 0 e 1, como podemos observar na figura 1, de tal forma que, com esta função podemos representar a probabilidade de que o evento de interesse ocorra. Note que a imagem da função $\boldsymbol{f(Z)}$ é o conjunto [0,1]. Assim, valores de $\boldsymbol{Z}$ são associados a valores de probabilidade.

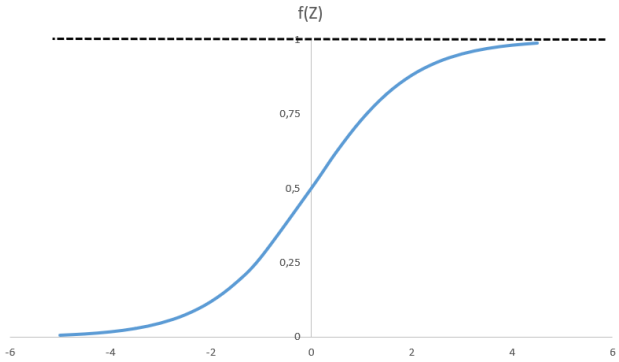


Figura 1: Gráfico da função logística com uma única variável.

Considerando $\boldsymbol{Z_i = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni}}$ onde $\boldsymbol{P}$ representa a probabilidade de ocorrência do evento de interesse, $\boldsymbol{x_i} = (\boldsymbol{x_{1i}, x_{2i}, \ldots, x_{ni}})$ representa o vetor das variáveis explicativas, no qual o subscrito $\mathbf{i}$ representa cada observação da amostra com $\mathbf{i = 1, \ldots, n}$ e $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_n}$ são os coeficientes a serem estimados do modelo. Isolando o valor da probabilidade em $\boldsymbol{f(Z) = \ln\left(\frac{P}{1-P}\right)}$ podemos escrever a probabilidade de ocorrência do evento sob estudo em função dos coeficientes estimados: $\boldsymbol{P = \frac{1}{1+e^{-f(Z)}}} =$

$\frac{1}{1+e^{-[\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n]}}$. O que a função logística faz é estimar a probabilidade de ocorrência do evento sob estudo em função da observação i.

Na equação apresentada, é importante verificar a presença de multicolinearidade entre as variáveis preditoras $(x_1, x_2; \ldots; x_n)$. A multicolinearidade pode afetar a confiabilidade dos coeficientes do modelo. Se houver forte correlação entre algumas das variáveis independentes, os coeficientes estimados podem ser pouco confiáveis. Para lidar com esse problema em regressões logísticas múltiplas, uma abordagem é usar a Análise Fatorial com o método de Análise de Componentes Principais.

Avaliação da qualidade do modelo de regressão logística

A estimação dos coeficientes de um modelo de regressão logística é realizada com a aplicação do método da máxima verossimilhança. Neste método busca-se maximizar a probabilidade de ocorrência de um evento.

De acordo com (HAIR, ANDERSON, TATHAM, BLACK, 2005), a medida, no caso da regressão logística que é similar a soma dos quadrados dos erros (resíduos) para a regressão linear é dada por -2 vezes o logaritmo do valor da verossimilhança e representada por -2LL ou -2log verossimilhança. Quanto menor for este valor, mais adequado é o modelo. Assim, se o valor da verossimilhança for igual a 1 (ajuste perfeito), então teremos que -2LL = 0. O valor da verossimilhança pode ser utilizado para comparações entre modelos, utilizando-se um modelo nulo sem as variáveis independentes que serve como referência para comparações com outros modelos propostos.

Existem muitas medidas que devem ser consideradas na verificação da qualidade dos modelos, tais como: Pseudo-R^2 Cox-Snell R^2, Nagelkerke R^2, Pseudo-R^2 de Tjur, Teste de hipótese qui-quadrado, teste de Razão de Verossimilhança (Likelihood ratio test), Deviance, Critério de informação de Akaike (AIC), Critério de informação Bayesiano (BIC), Teste Hosmer-Lemeshow, Teste da Significância dos coeficientes (Teste de Wald), Avaliação da capacidade de discriminação do modelo, Curva ROC e AUC e Variance Inflation Factor – VIF.

No entanto o interesse do presente artigo restringe-se apenas a questão da multicolinearidade, abordada estritamente no teste Variance Inflation Factor – VIF.

Análise de componentes principais (ACP)

Uma ferramenta da Estatística Multivariada para reduzir a multicolinearidade em um conjunto de variáveis é a Análise Fatorial com extração por Componentes Principais. Esta técnica permite que, a partir de um conjunto com N variáveis, em geral com elevada correlação entre parte destas variáveis, possamos obter um subconjunto de variáveis não correlacionadas entre si (chamadas de fatores) que são combinações lineares das variáveis originais e com uma relativa pequena perda da informação original dos dados. Dizer que teremos pequena perda de informação é equivalente a dizer que os fatores obtidos a partir das variáveis originais devem explicar o máximo da variância das dimensões da base de dados

original. Estas componentes principais são combinações lineares das variáveis originais.

Existem duas vantagens aqui: primeiro, eliminamos a correlação entre as variáveis originais, criando novas variáveis (chamadas de fatores) em termos das variáveis originais. Segundo, ao reduzir o conjunto de variáveis (agora chamadas de fatores), tornamos mais simples a tarefa de compreender o comportamento do fenômeno em estudo. Além disso, essa abordagem permite a descoberta de fatores subjacentes (variáveis latentes ou construtos) relacionados ao fenômeno.

Deve ser observado, ainda, que a primeira componente principal produzida pela ACP possui a maior variância possível da variância total dos dados (esta componente sozinha é a que consegue “explicar” a maior parte da variação dos dados). A segunda componente principal possui a segunda maior variância possível e assim por diante com todas as outras componentes principais.

Assim, a ACP permite um outro uso: o estudo de variáveis denominadas de variáveis latentes ou constructos (variáveis que não podem ser medidas diretamente, mas apenas indiretamente). A técnica ACP foi introduzida em 1901 por Karl Pearson e fundamentada por Hotelling em 1933 (MINGOTI, 2005). Os pesquisadores pioneiros da inteligência no início do século XX, Spearman e Thurstone fizeram uso da análise de fatores para entender a variável latente “inteligência” (FIELD, 2009). A ACP é usada atualmente para análise exploratória

de dados e na construção de modelos preditivos (LOESCH, HOELTGEBAUM, 2012).

Pressupostos da ACP

Para a aplicação da ACP, é necessário que duas condições principais sejam atendidas: (1) as variáveis devem ser métricas e (2) deve existir uma correlação elevada e estatisticamente significativa entre pelo menos um subconjunto de variáveis na base de dados.

Segundo (FÁVERO, BELFIORI, 2017), se, "com uma inspeção visual da matriz de correlações não indicar um número substancial de valores superior a 0,30, sua utilização provavelmente não será apropriada." O fato de termos conjuntos de coeficientes de correlação altos na matriz de correlação sugere que tais variáveis podem estar medindo uma mesma dimensão subjacente, ou seja, temos algum grau de redundância dentro destes conjuntos de variáveis correlacionadas.

Seguindo a notação de (FÁVERO, BELFIORI, SILVA, CHAN, 2009), considere as variáveis observáveis $(\boldsymbol{X_1, X_2, X_3, \dots X_p})$, extraídas de uma população de média $\boldsymbol{\mu} = \left(\boldsymbol{\mu_1, \mu_2, \dots \mu_p}\right)$ e matriz de covariância representada por $\boldsymbol{\Sigma}$. A partir da matriz de covariâncias podemos obter a matriz de correlações fazendo $\boldsymbol{\Lambda} = \frac{\boldsymbol{\Sigma}}{\sqrt{\boldsymbol{S_{ii}S_{jj}}}}$, onde $\boldsymbol{S_{ii}}$ representa as variâncias amostrais da i-ésima variável. O método das componentes principais pode utilizar tanto a matriz de covariâncias quanto a matriz de correlações. As componentes extraídas por um e outro método,

podem diferir, de forma geral. Recomenda-se utilizar a matriz de correlações, após mudanças de escalas de medida por estandardização (usando-se a média e o desvio padrão) para que se reduzam eventuais problemas de discrepância quando temos diferenças muito grandes entre as variâncias obtidas, causadas pelas diferentes unidades de medida (MINGOTI, 2005).

Assumimos que as variáveis observáveis são linearmente dependentes das variáveis não observáveis ($\boldsymbol{F_1, F_2, F_3, \dots F_m}$) que são os fatores comuns. Temos ainda $\boldsymbol{p}$ valores de variação específica associado a cada variável original representados por $\boldsymbol{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p}$. Escrevemos as variáveis observáveis $\boldsymbol{X_i}$ em função das variáveis não observáveis $\boldsymbol{F_j}$:

$$\boldsymbol{X_1 = \mu_1 + a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1}$$
$$\boldsymbol{X_2 = \mu_2 + a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2}$$
$$\dots$$
$$\boldsymbol{X_p = \mu_p + a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p}$$

Os coeficientes $\boldsymbol{a_{ij}}$ são denominados de loading ou cargas fatoriais. Eles representam o peso de cada variável $\boldsymbol{i}$ no fator $\boldsymbol{j}$ (o grau de correlação de Pearson entre as variáveis originais e os fatores). Após a padronização das variáveis originais cada variável acima ficará escrita como abaixo (FÁVERO, BELFIORI, SILVA, CHAN, 2009):

$$\boldsymbol{X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i}$$

Os fatores $\boldsymbol{F_j}$ são perpendiculares (independentes) entre si, com média zero e variância igual a 1 enquanto os fatores específicos $\boldsymbol{\varepsilon_i}$ possuem média zero e variância $\boldsymbol{\psi_i, i = 1, \dots, p}$.

Para a extração das componentes principais precisamos de conhecimentos de álgebra linear. Serão extraídos os autovalores e autovetores da matriz de correlações padronizadas. A partir das variáveis originais $\boldsymbol{X_i}$ obtemos as variáveis padronizadas $\boldsymbol{Z_i = \frac{X_i - \mu_i}{\sigma_i}}$ onde $\boldsymbol{E(X_i) = \mu_i}$ **e** $\boldsymbol{Var(X_i) = \sigma_i^2}$. Construímos a matriz de covariâncias das variáveis $\boldsymbol{Z_i}$ representada por $\boldsymbol{P_{p \times p}}$.

Considere uma matriz ($\boldsymbol{p \times p}$) na qual $\boldsymbol{p}$ é o número de variáveis observadas. A matriz será representada por: $\boldsymbol{\Lambda} = \begin{bmatrix} 1 & \rho_{1\times2}\rho_{1\times3}\cdots\rho_{1\times p} \\ \rho_{1\times2} & 1 & \rho_{3\times2}\cdots\rho_{2\times p} \\ \dots & \dots & \dots \dots \dots \\ \rho_{1\times p}\rho_{2\times p}\rho_{3\times p}\cdots & 1 \end{bmatrix}_{p\times p}$, a correlação da variável i com a variável j é igual a correlação da variável j com a variável i, donde a matriz de correlação é simétrica. Da álgebra linear sabemos que toda matriz real e simétrica é positiva definida. Portanto, todos os seus autovalores são números reais positivos.

Contudo, antes de prosseguir com a ACP, é necessário verificar se faz sentido utilizar a técnica. Se não existir correlação forte e estatisticamente significativa entre subconjuntos de variáveis dentre as variáveis observadas, não devemos prosseguir com a ACP. Com a finalidade de testar a adequação da ACP utiliza-se o teste de esfericidade de Bartlett. A hipótese que se está testando aqui é: os coeficientes de correlação são estatisticamente distintos de zero? Em outras

palavras, estamos testando se a matriz de correlações é estatisticamente distinta da matriz identidade. Em símbolos teremos:

$$\boldsymbol{H_0}: \boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\rho_{1\times2}} & \boldsymbol{\rho_{1\times3}} \cdots & \boldsymbol{\rho_{1\times p}} \\ \boldsymbol{\rho_{1\times2}} & \mathbf{1} & \boldsymbol{\rho_{3\times2}} \cdots & \boldsymbol{\rho_{2\times p}} \\ \cdots & \cdots & \cdots \cdots & \cdots \\ \boldsymbol{\rho_{1\times p}} & \boldsymbol{\rho_{2\times p}} & \boldsymbol{\rho_{3\times p}} \cdots & \mathbf{1} \end{bmatrix}_{p\times p} = \boldsymbol{I} = \begin{bmatrix} \mathbf{1\,0\,0} \cdots \mathbf{0} \\ \mathbf{0\,1\,0} \cdots \mathbf{0} \\ \cdots\cdots\cdots\cdots \\ \mathbf{0\,0\,0} \cdots \mathbf{1} \end{bmatrix}_{p\times p}$$

$$\boldsymbol{H_1}: \boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\rho_{1\times2}} & \boldsymbol{\rho_{1\times3}} \cdots & \boldsymbol{\rho_{1\times p}} \\ \boldsymbol{\rho_{1\times2}} & \mathbf{1} & \boldsymbol{\rho_{3\times2}} \cdots & \boldsymbol{\rho_{2\times p}} \\ \cdots & \cdots & \cdots \cdots & \cdots \\ \boldsymbol{\rho_{1\times p}} & \boldsymbol{\rho_{2\times p}} & \boldsymbol{\rho_{3\times p}} \cdots & \mathbf{1} \end{bmatrix}_{p\times p} \neq \boldsymbol{I} = \begin{bmatrix} \mathbf{1\,0\,0} \cdots \mathbf{0} \\ \mathbf{0\,1\,0} \cdots \mathbf{0} \\ \cdots\cdots\cdots\cdots \\ \mathbf{0\,0\,0} \cdots \mathbf{1} \end{bmatrix}_{p\times p}$$

$\boldsymbol{\chi^2_{Bartlett}} = -\left[(\boldsymbol{n}-\mathbf{1}) - \left(\frac{\mathbf{2}\boldsymbol{p}+\mathbf{5}}{\mathbf{6}}\right)\right]\mathbf{ln}\ |\boldsymbol{D}|$, com $\frac{\boldsymbol{p}(\boldsymbol{p}-\mathbf{1})}{\mathbf{2}}$ graus de liberdade, n o tamanho da amostra e $\boldsymbol{D}$ é o determinante da matriz de correlação.

Caso a hipótese nula não seja rejeitada, deve-se desconsiderar a utilização da ACP (as variáveis não são correlacionadas).

Uma estatística recomendada para se verificar a adequabilidade do uso da ACP é a de Kaiser-Meyer-Olkin (KMO). Esta estatística compara correlações simples com correlações parciais e é dada por:

$$\boldsymbol{KMO} = \frac{\sum_{i\neq j}\sum \boldsymbol{r_{ij}^2}}{\sum_{i\neq j}\sum \boldsymbol{r_{ij}^2} + \sum_{i\neq j}\sum \boldsymbol{\varphi_{ij}^2}}$$

Onde:

$\boldsymbol{r_{ij}}$: coeficiente de correlação entre as variáveis

$\boldsymbol{\varphi_{ij}}$: coeficiente de correlação parcial

Quanto mais próxima de 1,0 for a estatística KMO, mais adequada é a ACP. Valores próximos de 0,0 indicam que a ACP não é adequada, indicando correlação fraca entre as variáveis. A tabela a seguir apresenta os valores da estatística KMO e o quanto a Análise Fatorial é indicada em cada caso.

KMO	Análise Fatorial
0,9 – 1,0	Muito boa
0,8 – 0,9	Boa
0,7 – 0,8	Média
0,6 – 0,7	Razoável
0,5 – 0,6	Má
< 0,5	Inaceitável

Tabela 2: Fonte: (FÁVERO, BELFIORI, SILVA, CHAN, 2009).

Matriz anti-imagem

De acordo com (MAROCO, 2007) as matrize de anti-imagem para variância e covariância e para as correlações "apresentam os valores negativos das covariâncias e correlações parciais entre as variáveis. Estes valores estimam as correlações entre as variáveis que não são devidas aos fatores comuns. Valores baixos destas correlações parciais indicam que as variáveis partilham um ou mais fatores comuns, enquanto valores altos sugerem que as variáveis são mais ou menos independentes. Assim, os valores abaixo da diagonal devem ser próximos de zero". (MAROCO, 2007) recomenda que os valores da diagonal principal da matriz anti-imagem que são menores que 0,5 devem ser desconsiderados da Análise Fatorial.

(FÁVERO, BELFIORI, 2009) destacam que o teste de esfericidade de Bartlett deve ser sempre preferido em relação à estatística KMO, já que o primeiro é um

teste com nível de significância determinado, enquanto o segundo é apenas um coeficiente, sem uma distribuição de probabilidade determinada.

Tendo obtido os autovalores e autovetores, podemos determinar os escores fatoriais.

Denotando a matriz identidade por $\boldsymbol{I}$, os autovalores $\boldsymbol{\lambda_i^2}$ da matriz $\boldsymbol{\Lambda}$ são obtidos resolvendo-se a equação: $\mathbf{det}(\boldsymbol{\Lambda} - \boldsymbol{\lambda}\mathbf{I}) = \mathbf{0}$.

Resolvendo a equação temos p autovetores $\boldsymbol{\lambda^2}$ ($\boldsymbol{\lambda_1^2 \geq \lambda_2^2 \geq \lambda_3^2 \geq \cdots \geq \lambda_p^2}$).

Além disso, temos que $\boldsymbol{\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_k^2 = k}$.

Para cada autovalor calcula-se o respectivo autovetor resolvendo-se a equação

$$(\boldsymbol{\Lambda} - \boldsymbol{\lambda^2}\mathbf{I})\mathbf{v} = \mathbf{0}$$

Tendo obtidos os autovalores e os autovetores, temos como obter os escores fatoriais aqui representados por $\boldsymbol{s_1, s_2, \ldots, s_k}$. Os escores fatoriais são os valores numéricos que relacionam as variáveis originais com os fatores.

Os escores são dados pelas expressões a seguir:

$$\boldsymbol{s_1} = \begin{bmatrix} \boldsymbol{s_{11}} \\ \boldsymbol{s_{21}} \\ \cdots \\ \boldsymbol{s_{k1}} \end{bmatrix} = \begin{bmatrix} \dfrac{\boldsymbol{v_{11}}}{\sqrt{\boldsymbol{\lambda_1^2}}} \\ \dfrac{\boldsymbol{v_{21}}}{\sqrt{\boldsymbol{\lambda_1^2}}} \\ \cdots \\ \dfrac{\boldsymbol{v_{k1}}}{\sqrt{\boldsymbol{\lambda_1^2}}} \end{bmatrix}, \boldsymbol{s_1} = \begin{bmatrix} \boldsymbol{s_{12}} \\ \boldsymbol{s_{22}} \\ \cdots \\ \boldsymbol{s_{k2}} \end{bmatrix} = \begin{bmatrix} \dfrac{\boldsymbol{v_{12}}}{\sqrt{\boldsymbol{\lambda_2^2}}} \\ \dfrac{\boldsymbol{v_{22}}}{\sqrt{\boldsymbol{\lambda_2^2}}} \\ \cdots \\ \dfrac{\boldsymbol{v_{k2}}}{\sqrt{\boldsymbol{\lambda_2^2}}} \end{bmatrix}, \ldots, \boldsymbol{s_1} = \begin{bmatrix} \boldsymbol{s_{1k}} \\ \boldsymbol{s_{2k}} \\ \cdots \\ \boldsymbol{s_{kk}} \end{bmatrix} = \begin{bmatrix} \dfrac{\boldsymbol{v_{1k}}}{\sqrt{\boldsymbol{\lambda_k^2}}} \\ \dfrac{\boldsymbol{v_{2k}}}{\sqrt{\boldsymbol{\lambda_k^2}}} \\ \cdots \\ \dfrac{\boldsymbol{v_{kk}}}{\sqrt{\boldsymbol{\lambda_k^2}}} \end{bmatrix},$$

Tendo os escores fatoriais obtemos os fatores, a partir das variáveis originais transformadas pelo Z-score (aqui representadas por $\boldsymbol{ZX_{ij}}$):

$$\boldsymbol{F_{1i} = \frac{v_{11}}{\sqrt{\lambda_1^2}} ZX_{1i} + \frac{v_{21}}{\sqrt{\lambda_1^2}} ZX_{2i} + \cdots + \frac{v_{k1}}{\sqrt{\lambda_1^2}} ZX_{ki}}$$

$$\boldsymbol{F_{2i} = \frac{v_{12}}{\sqrt{\lambda_2^2}} ZX_{1i} + \frac{v_{22}}{\sqrt{\lambda_2^2}} ZX_{2i} + \cdots + \frac{v_{k2}}{\sqrt{\lambda_2^2}} ZX_{ki}}$$

...

$$\boldsymbol{F_{ki} = \frac{v_{1k}}{\sqrt{\lambda_k^2}} ZX_{1i} + \frac{v_{2k}}{\sqrt{\lambda_k^2}} ZX_{2i} + \cdots + \frac{v_{kk}}{\sqrt{\lambda_k^2}} ZX_{ki}}$$

Agora que temos os fatores, é necessário definir quantos fatores devem ser retidos. O critério de Kaiser sugere que devam ser mantidos aqueles fatores associados aos autovalores maiores que 1.

Os autovalores representam a porcentagem da variância compartilhada pelas variáveis originais na constituição de cada fator. Assim, no caso de se extrair um fator a partir de algum autovalor com valor menor que 1, este fator dificilmente conseguirá representar o comportamento de variáveis originais. Pelo critério da raiz latente só devem ser mantidos os fatores associados com autovalores maiores que 1.

Para cada variável original $\boldsymbol{X_i}$ temos que sua variância pode ser decomposta em duas partes: $\boldsymbol{Var(X_i) = a_{i1}^2 + a_{i2}^2 + a_{i3}^2 \ldots + a_{ip}^2 + \psi_i}$

Chama-se de comunalidade à expressão $\boldsymbol{h_i^2 = a_{i1}^2 + a_{i2}^2 + a_{i3}^2 \ldots + a_{ip}^2}$

E variância específica é a grandeza $\boldsymbol{\psi_i}$ e não está relacionada aos fatores comuns.

A comunalidade é uma estimativa da variância de $\boldsymbol{X_i}$ que é explicada pelos fatores comuns, sendo a variância total que é compartilhada com uma variável original com as outras variáveis. Valores muito baixos de comunalidade sugerem que o pesquisador retire a correspondente variável da análise fatorial.

A Variância da primeira componente principal é a maior de todas. A soma das variâncias se mantém constante e igual a 1: $\boldsymbol{\gamma_{i1}^2 + \gamma_{i2}^2 + \gamma_{i3}^2 + \cdots + \gamma_{ip}^2 = 1}$.

A comunalidade: $\boldsymbol{h_i^2 = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2}$ representa uma estimativa da variância das variáveis originais $\boldsymbol{X_i}$, explicada pelos fatores comuns.

Quantos fatores devem ser retidos?

Uma das principais discussões na ACP é quantos fatores devem ser retidos de tal forma a não ocorrer perda significativa da informação contida nas variáveis originais? Aqui vamos adotar o critério da raiz latente (também denominado critério de Kaiser). Lembremos que os autovalores são associados à porcentagem da variância compartilhada pelas variáveis originais na composição de cada fator. O fator associado com o maior autovalor é o que explica a maior parte dessa variância. Assim, escolhemos fatores associados apenas com autovetores maiores que 1 já que fatores que sejam extraídos de autovalores menores que 1 não conseguem explicar nem ao menos uma das variáveis originais.

Outra abordagem para definir quantos fatores devem ser retidos é o Gráfico Scree. Neste gráfico cada uma das componentes principais é representada no eixo x e os autovalores associados no eixo y. Por análise visual deste gráfico devemos reter todas as componentes até que o gráfico fique horizontal, significando que a contribuição à variância total da próxima componente já não é tão relevante. Claramente, a principal crítica a este critério é sua subjetividade.

Rotação dos fatores

Realizada a extração dos fatores, podemos melhorara a distribuição das cargas fatoriais por meio do procedimento denominado de rotação dos fatores. A rotação dos fatores constitui alternativa para maximizar as cargas das variáveis originais em cada fator, enquanto diminui as cargas em outras variáveis, assim, por vezes, permitindo uma redistribuição das cargas fatoriais e permitindo, eventualmente, reduzir a quantidade de variáveis com elevadas cargas para um certo fator e podendo simplificar a interpretação dos fatores. Importante (FÁVERO, BELFIORI, 2017): as rotações não modifiquem as comunalidades, o percentual total de variância compartilhada pelas variáveis nos fatores, nem a estatística KMO nem o qui-quadrado de Bartlett, temos uma redistribuição da porcentagem de variância compartilhada pelas variáveis originais entre os fatores. Assim, são determinados novos autovalores a partir da rotação. Também são obtidos novos escores fatoriais rotacionados. Mas, claro, a matriz de correlação não é modificada pela rotação. Existem vários procedimentos de rotação: os ortogonais e os oblíquos. Os principais tipos de rotação ortogonal

são: Varimax, Quartimax e equimax. Já os principais procedimentos de rotação oblíqua são: oblimin, quartimin e promax. Fonte: https://www.blog.psicometriaonline.com.br/rotacao-fatorial/ acesso em 04/07/2024.

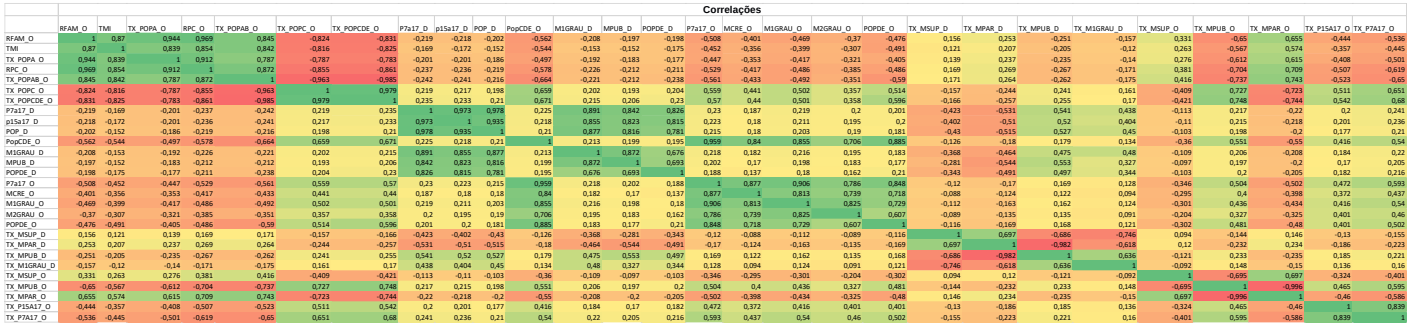
Neste artigo adotamos a técnica Varimax. Este tipo de rotação minimiza a quantidade de variáveis com carga elevada em certo fator redistribuindo as cargas fatoriais e maximizando a variância compartilhada nos fatores associados com autovalores de menor valor.

A seguir apresentamos a técnica ACP para redução das variáveis originais, na sequência, aplicamos os fatores obtidos a uma regressão logística para modelar a divisão modal a partir dos dados da Pesquisa OD 2017.

Aplicando ACP aos dados da Pesquisa OD 2017

Tal como explicado anteriormente, o primeiro passo para aplicar a ACP é verificar a matriz de correlações entre as variáveis do estudo. Caso não tenhamos correlações fortes (positivas ou negativas entre subconjuntos de variáveis, a ACP não é recomendada). Esta matriz está apresentada na figura 2 a seguir. O verde mais escuro

indica correlações positivas fortes e o vermelho mais escuro indica correlações negativas fortes.

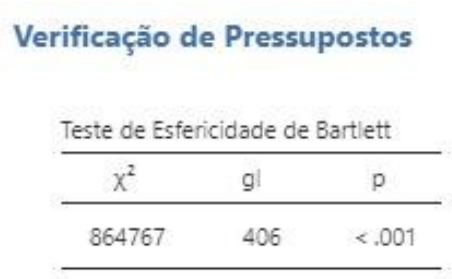


Correlações

	RFAM_O	TMI	TX_POPA_O	RPC_O	TX_POPAB_O	TX_POPC_O	TX_POPCDE_O	P7a17_D	p15a17_D	POP_D	PopCDE_O	M1GRAU_D	MPUB_D	POPDE_D	P7a17_O	MCRE_O	M1GRAU_O	M2GRAU_O	POPDE_O	TX_MSUP_D	TX_MPAR_D	TX_MPUB_D	TX_M1GRAU_D	TX_MSUP_O	TX_MPUB_O	TX_MPAR_O	TX_P15A17_O	TX_P7A17_O
RFAM_O	1	0,87	0,944	0,969	0,845	-0,824	-0,831	-0,219	-0,218	-0,202	-0,562	-0,208	-0,197	-0,198	-0,508	-0,401	-0,469	-0,37	-0,476	0,156	0,253	-0,251	-0,157	0,331	-0,65	0,655	-0,444	-0,536
TMI	0,87	1	0,839	0,854	0,842	-0,816	-0,825	-0,169	-0,172	-0,152	-0,544	-0,153	-0,152	-0,175	-0,452	-0,356	-0,399	-0,307	-0,491	0,121	0,207	-0,205	-0,12	0,263	-0,567	0,574	-0,357	-0,445
TX_POPA_O	0,944	0,839	1	0,912	0,787	-0,787	-0,783	-0,201	-0,201	-0,186	-0,497	-0,192	-0,183	-0,177	-0,447	-0,353	-0,417	-0,321	-0,405	0,139	0,237	-0,235	-0,14	0,276	-0,612	0,615	-0,408	-0,501
RPC_O	0,969	0,854	0,912	1	0,872	-0,855	-0,861	-0,237	-0,236	-0,219	-0,578	-0,226	-0,212	-0,211	-0,529	-0,417	-0,486	-0,385	-0,486	0,169	0,269	-0,267	-0,171	0,381	-0,704	0,709	-0,507	-0,619
TX_POPAB_O	0,845	0,842	0,787	0,872	1	-0,963	-0,985	-0,242	-0,241	-0,216	-0,664	-0,221	-0,212	-0,238	-0,561	-0,433	-0,492	-0,351	-0,59	0,171	0,264	-0,262	-0,175	0,416	-0,737	0,743	-0,523	-0,65
TX_POPC_O	-0,824	-0,816	-0,787	-0,855	-0,963	1	0,979	0,219	0,217	0,198	0,659	0,202	0,193	0,204	0,559	0,441	0,502	0,357	0,514	-0,157	-0,244	0,241	0,161	-0,409	0,727	-0,723	0,511	0,651
TX_POPCDE_O	-0,831	-0,825	-0,783	-0,861	-0,985	0,979	1	0,235	0,233	0,21	0,671	0,215	0,206	0,23	0,57	0,44	0,501	0,358	0,596	-0,166	-0,257	0,255	0,17	-0,421	0,748	-0,744	0,542	0,68
P7a17_D	-0,219	-0,169	-0,201	-0,237	-0,242	0,219	0,235	1	0,973	0,978	0,225	0,891	0,842	0,826	0,23	0,187	0,219	0,2	0,201	-0,423	-0,531	0,541	0,438	-0,113	0,217	-0,22	0,2	0,241
p15a17_D	-0,218	-0,172	-0,201	-0,236	-0,241	0,217	0,233	0,973	1	0,935	0,218	0,855	0,823	0,815	0,223	0,18	0,211	0,195	0,2	-0,402	-0,51	0,52	0,404	-0,11	0,215	-0,218	0,201	0,236
POP_D	-0,202	-0,152	-0,186	-0,219	-0,216	0,198	0,21	0,978	0,935	1	0,21	0,877	0,816	0,781	0,215	0,18	0,203	0,19	0,181	-0,43	-0,515	0,527	0,45	-0,103	0,198	-0,2	0,177	0,21
PopCDE_O	-0,562	-0,544	-0,497	-0,578	-0,664	0,659	0,671	0,225	0,218	0,21	1	0,213	0,199	0,195	0,959	0,84	0,855	0,706	0,885	-0,126	-0,18	0,179	0,134	-0,36	0,551	-0,55	0,416	0,54
M1GRAU_D	-0,208	-0,153	-0,192	-0,226	-0,221	0,202	0,215	0,891	0,855	0,877	0,213	1	0,872	0,676	0,218	0,182	0,216	0,195	0,183	-0,368	-0,464	0,475	0,48	-0,109	0,206	-0,208	0,184	0,22
MPUB_D	-0,197	-0,152	-0,183	-0,212	-0,212	0,193	0,206	0,842	0,823	0,816	0,199	0,872	1	0,693	0,202	0,17	0,198	0,183	0,177	-0,281	-0,544	0,553	0,327	-0,097	0,197	-0,2	0,17	0,205
POPDE_D	-0,198	-0,175	-0,177	-0,211	-0,238	0,204	0,23	0,826	0,815	0,781	0,195	0,676	0,693	1	0,188	0,137	0,18	0,162	0,21	-0,343	-0,491	0,497	0,344	-0,103	0,2	-0,205	0,182	0,216
P7a17_O	-0,508	-0,452	-0,447	-0,529	-0,561	0,559	0,57	0,23	0,223	0,215	0,959	0,218	0,202	0,188	1	0,877	0,906	0,786	0,848	-0,12	-0,17	0,169	0,128	-0,346	0,504	-0,502	0,472	0,593
MCRE_O	-0,401	-0,356	-0,353	-0,417	-0,433	0,441	0,44	0,187	0,18	0,18	0,84	0,182	0,17	0,137	0,877	1	0,813	0,739	0,718	-0,088	-0,124	0,122	0,094	-0,295	0,4	-0,398	0,372	0,437
M1GRAU_O	-0,469	-0,399	-0,417	-0,486	-0,492	0,502	0,501	0,219	0,211	0,203	0,855	0,216	0,198	0,18	0,906	0,813	1	0,825	0,729	-0,112	-0,163	0,162	0,124	-0,301	0,436	-0,434	0,416	0,54
M2GRAU_O	-0,37	-0,307	-0,321	-0,385	-0,351	0,357	0,358	0,2	0,195	0,19	0,706	0,195	0,183	0,162	0,786	0,739	0,825	1	0,607	-0,089	-0,135	0,135	0,091	-0,204	0,327	-0,325	0,401	0,46
POPDE_O	-0,476	-0,491	-0,405	-0,486	-0,59	0,514	0,596	0,201	0,2	0,181	0,885	0,183	0,177	0,21	0,848	0,718	0,729	0,607	1	-0,116	-0,169	0,168	0,121	-0,302	0,481	-0,48	0,401	0,502
TX_MSUP_D	0,156	0,121	0,139	0,169	0,171	-0,157	-0,166	-0,423	-0,402	-0,43	-0,126	-0,368	-0,281	-0,343	-0,12	-0,088	-0,112	-0,089	-0,116	1	0,697	-0,686	-0,746	0,094	-0,144	0,146	-0,13	-0,155
TX_MPAR_D	0,253	0,207	0,237	0,269	0,264	-0,244	-0,257	-0,531	-0,51	-0,515	-0,18	-0,464	-0,544	-0,491	-0,17	-0,124	-0,163	-0,135	-0,169	0,697	1	-0,982	-0,618	0,12	-0,232	0,234	-0,186	-0,223
TX_MPUB_D	-0,251	-0,205	-0,235	-0,267	-0,262	0,241	0,255	0,541	0,52	0,527	0,179	0,475	0,553	0,497	0,169	0,122	0,162	0,135	0,168	-0,686	-0,982	1	0,636	-0,121	0,233	-0,235	0,185	0,221
TX_M1GRAU_D	-0,157	-0,12	-0,14	-0,171	-0,175	0,161	0,17	0,438	0,404	0,45	0,134	0,48	0,327	0,344	0,128	0,094	0,124	0,091	0,121	-0,746	-0,618	0,636	1	-0,092	0,148	-0,15	0,136	0,16
TX_MSUP_O	0,331	0,263	0,276	0,381	0,416	-0,409	-0,421	-0,113	-0,11	-0,103	-0,36	-0,109	-0,097	-0,103	-0,346	-0,295	-0,301	-0,204	-0,302	0,094	0,12	-0,121	-0,092	1	-0,695	0,697	-0,324	-0,401
TX_MPUB_O	-0,65	-0,567	-0,612	-0,704	-0,737	0,727	0,748	0,217	0,215	0,198	0,551	0,206	0,197	0,2	0,504	0,4	0,436	0,327	0,481	-0,144	-0,232	0,233	0,148	-0,695	1	-0,996	0,465	0,595
TX_MPAR_O	0,655	0,574	0,615	0,709	0,743	-0,723	-0,744	-0,22	-0,218	-0,2	-0,55	-0,208	-0,2	-0,205	-0,502	-0,398	-0,434	-0,325	-0,48	0,146	0,234	-0,235	-0,15	0,697	-0,996	1	-0,46	-0,586
TX_P15A17_O	-0,444	-0,357	-0,408	-0,507	-0,523	0,511	0,542	0,2	0,201	0,177	0,416	0,184	0,17	0,182	0,472	0,372	0,416	0,401	0,401	-0,13	-0,186	0,185	0,136	-0,324	0,465	-0,46	1	0,839
TX_P7A17_O	-0,536	-0,445	-0,501	-0,619	-0,65	0,651	0,68	0,241	0,236	0,21	0,54	0,22	0,205	0,216	0,593	0,437	0,54	0,46	0,502	-0,155	-0,223	0,221	0,16	-0,401	0,595	-0,586	0,839	1

Figura 2: matriz de correlações entre as variáveis em estudo. Fonte: os autores.

O passo seguinte é efetuar o teste de esfericidade de Bartlett. Como o p-valor é menor que o nível de significância adotado de 5%, rejeitamos a H0 e concluímos que a matriz de correlação é estatisticamente diferente da matriz unidade.



Verificação de Pressupostos

Teste de Esfericidade de Bartlett

χ^2	gl	p
864767	406	< .001

Figura 3: Teste de esfericidade de Bartlett.

Continuando ainda com a verificação de pressupostos, analisamos agora a estatística KMO. Como vimos, quanto mais próxima de 1,0 estiver esta estatística, mais adequada é a aplicação da ACP. Basicamente os valores da estatística KMO se situam entre Boa e Muito Boa, de acordo com a Tabela. 2.

Medida de Adequação de Amostragem de KMO	
	MAA
Global	0.889
16_TMI	0.979
6_RPC_O	0.927
7_RFAM_O	0.889
12_M1GRAU_O	0.954
13_M2GRAU_O	0.944
10_MCRE_O	0.973
1_TX_MPUB_O	0.825
2_TX_MPAR_O	0.825
17_TX_MSUP_O	0.982
5_POPDE_O	0.848
14_PopCDE_O	0.879
18_P7a17_O	0.889
15_TX_POPA_O	0.945
19_TX_POPC_O	0.873
20_TX_POPAB_O	0.883

Medida de Adequação de Amostragem de KMO	
	MAA
21_TX_POPCDE_O	0.846
9_TX_P7A17_O	0.865
26_TX_P15A17_O	0.897
11_M1GRAU_D	0.892
27_MPUB_D	0.913
3_TX_MPUB_D	0.823
4_TX_MPAR_D	0.821
22_TX_M1GRAU_D	0.834
28_TX_MSUP_D	0.888
23_POP_D	0.906
8_POPDE_D	0.920
24_PopCDE_D	0.924
25_P7a17_D	0.849
29_p15a17_D	0.902

Figura 4: Estatística KMO. Fonte: os autores.

Na sequência calculamos os autovalores. Na figura 5 vemos a porcentagem da variância total associada a cada autovalor e a porcentagem acumulada.

Note que a variância acumulada explicada pelas cinco primeiras componentes corresponde a 83,9% do total da variância. Assim, reduzimos de um total de 29 variáveis para 5 componentes principais, perdendo no processo, cerca de 16,1% da variância.

Valores próprios

Valores próprios iniciais

Componente	Valor próprio	% de Variância total	% acumulada
1	12.47416	43.01436	43.0
2	6.06978	20.93028	63.9
3	2.66138	9.17718	73.1
4	1.78005	6.13810	79.3
5	1.34992	4.65489	83.9
6	0.98877	3.40955	87.3
7	0.59511	2.05210	89.4
8	0.55254	1.90530	91.3
9	0.42026	1.44917	92.7
10	0.31148	1.07408	93.8
11	0.29225	1.00776	94.8
12	0.25629	0.88376	95.7
13	0.22494	0.77567	96.5
14	0.18988	0.65475	97.1
15	0.16146	0.55677	97.7

Valores próprios

Valores próprios iniciais

Componente	Valor próprio	% de Variância total	% acumulada
16	0.12006	0.41401	98.1
17	0.11632	0.40110	98.5
18	0.08027	0.27679	98.8
19	0.07981	0.27521	99.1
20	0.07534	0.25979	99.3
21	0.07175	0.24741	99.6
22	0.03596	0.12400	99.7
23	0.02380	0.08208	99.8
24	0.01988	0.06853	99.8
25	0.01708	0.05891	99.9
26	0.01385	0.04774	99.9
27	0.00888	0.03061	100.0
28	0.00668	0.02303	100.0
29	0.00205	0.00707	100.0

Figura 5: Autovalores, com respectiva porcentagem e porcentagem acumulada. Fonte: os autores.

O próximo passo é definir quantas componentes devem ser retidas. Pelo critério de Kaiser, devem ser retidas aquelas componentes associadas com autovalores cujo valor é maior que 1. Com isso, serão retidas as cinco primeiras componentes. Alternativamente, temos o critério do gráfico Scree, apresentado na figura 6. Por ele também sugere-se reter as cinco primeiras componentes.

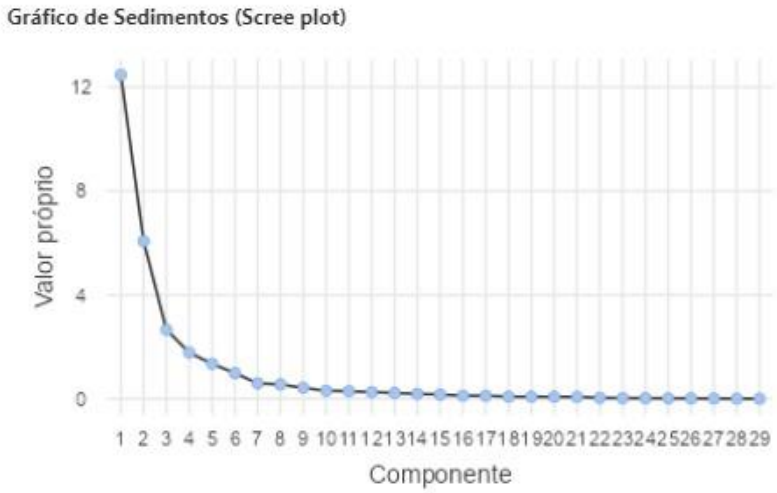


Figura 6: Gráfico Scree. Fonte: os autores.

Na figura 7 vemos estatísticas das correlações entre estas 5 componentes retidas. Como esperado, as correlações entre as componentes principais são nulas. Assim, está resolvido o problema da multicolinearidade e podemos aplicar estas componentes principais na Regressão Logística Múltipla.

Estatísticas das Componentes

Correlações Inter-componentes

	1	2	3	4	5	6
1	—	0.00	0.00	0.00	0.00	0.00
2		—	0.00	0.00	0.00	0.00
3			—	0.00	0.00	0.00
4				—	0.00	0.00
5					—	0.00
6						—

Figura 7: Correlações inter-componentes principais. Fonte: os autores.

Aplicando Regressão Logística aos fatores principais

Agora aplicamos as componentes principais para desenvolver um modelo de regressão logística binomial, cuja variável resposta é o modo de transporte assumindo os valores TI = Transporte Individual e TC = Transporte Coletivo. Lembrando que estamos tratando neste texto do modelo de divisão modal para o motivo Base Domiciliar Trabalho (BDT). Os coeficientes do modelo estão apresentados na figura 8 a seguir. Observe que o p-valor para todos os componentes principais foram significativos. Além disso, os valores do VIF (Valor de Inflação da Variância) foram todos abaixo de 10, confirmando que não temos multicolinearidade nas novas variáveis.

Coeficientes do modelo - TIPVG_TTC

		Intervalo de Confiança a 95%						Intervalo de Confiança a 95%	
Preditor	Estimativas	Lim. Inferior	Superior	Erro-padrão	Z	p	Rácio das Chances	Lim. Inferior	Superior
Intercepto	−0.0113	−0.0446	0.0221	0.0170	−0.661	0.508	0.989	0.956	1.022
CG_TC	0.3477	0.2963	0.3992	0.0263	13.245	< .001	1.416	1.345	1.491
CG_TI	−0.5680	−0.6259	−0.5102	0.0295	−19.244	< .001	0.567	0.535	0.600
Pontuação Compoenente 1	0.5302	0.4936	0.5668	0.0187	28.390	< .001	1.699	1.638	1.763
Pontuação Compoenente 2	0.0806	0.0458	0.1154	0.0178	4.542	< .001	1.084	1.047	1.122
Pontuação Compoenente 3	−0.0977	−0.1322	−0.0633	0.0176	−5.553	< .001	0.907	0.876	0.939
Pontuação Compoenente 4	0.1224	0.0874	0.1573	0.0178	6.864	< .001	1.130	1.091	1.170

Nota. As estimativas representam o Log das Chances de "TIPVG_TTC = TI" vs. "TIPVG_TTC = TC"

Verificação de Pressupostos

Segundo Andy Field pág.914 - Valores de tolerância inferiores a 0,1 e VIF maiores do que 10 são preocupantes. Observando-se a tale a seguir, todas as variáveis são aprovadas no pressuposto de não multicolinearidade.

Estatísticas de Colinearidade

	VIF	Tolerância
CG_TC	2.35	0.425
CG_TI	2.54	0.394
Pontuação Compoenente 1	1.07	0.935
Pontuação Compoenente 2	1.12	0.889
Pontuação Compoenente 3	1.05	0.955
Pontuação Compoenente 4	1.09	0.917

[4]

Figura 8: coeficientes do modelo de regressão logística. Fonte: os autores.

Na figura 9 vemos as medidas de ajuste do modelo de regressão logística, com particular destaque para o pseudo-R^2 de Cox-Snell. O teste omnibus de razão de verossimilhança mostra todas as componentes principais como significativas.

Regressão logística binomial

Medidas de Ajustamento do Modelo

			Teste ao Modelo Global		
Modelo	Desviância	R^2_{CS}	χ^2	gl	p
1	20010	0.162	2068	6	< .001
2	19917	0.169	2161	7	< .001

Comparações de Modelos

Comparação					
Modelo		Modelo	χ^2	gl	p
1	-	2	92.9	1	< .001

Resultados específicos do modelo Modelo 1

Esse é o melhor modelo com as variáveis disponíveis no banco BDTS_selecionadas_clusters

Teste omnibus do rácio de verosimilhanças

Preditor	χ^2	gl	p
CG_TC	183.0	1	< .001
CG_TI	403.5	1	< .001
Pontuação Compoenente 1	885.1	1	< .001
Pontuação Compoenente 2	20.7	1	< .001
Pontuação Compoenente 3	31.0	1	< .001
Pontuação Compoenente 4	47.3	1	< .001

[4]

Figura 9: Medidas de ajuste do modelo de Regressão Logística. Fonte: os autores.

Na figura 10 vemos um gráfico mostrando o limiar adotado (cutoff ou treshold) para fins de classificação pelo modelo logístico. Adotamos limiar = 0,5.

Figura 10: linhas de cutoff (threshold) entre sensibilidade e especificidade. Fonte: os autores.

Na figura 11 vemos a tabela de classificação (ou tabela de confusão). E as medidas de qualidade de ajuste do modelo logo abaixo.

Tabela de Classificação – ...

	Previsto		
Observado	TC	TI	% Correto
TC	4996	2977	62.7
TI	2647	5306	66.7

Nota. O valor de corte é 0.5

Medidas Preditivas

Acurácia	Especificidade	Sensibilidade	AUC
0.647	0.627	0.667	0.702

Nota. O valor de corte é 0.5

Figura 11: matriz de confusão e medidas do modelo logístico. Fonte: os autores.

Na figura 12 apresentamos a curva ROC. Como observado anteriormente, quanto maior a Área sob a curva (AUC) melhor o modelo de classificação.

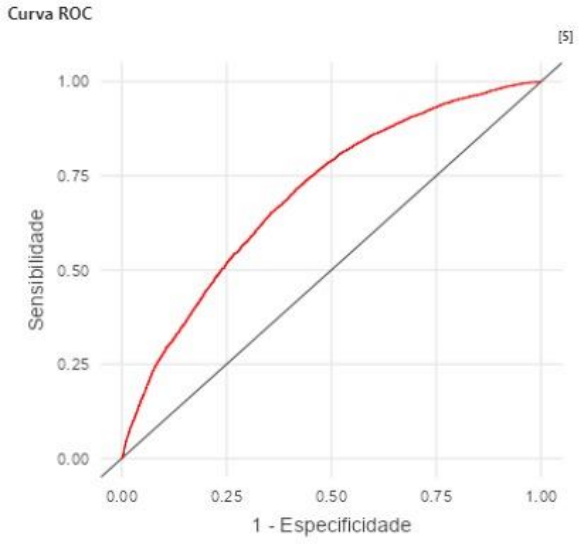


Figura 12: Curva ROC. Fonte: os autores.

Finalmente, queremos comparar as amplitudes dos intervalos de confiança obtidas neste modelo via ACP e em um modelo aplicando as variáveis sem ACP. Os intervalos de confiança para o modelo sem ACP estão na figura 13.

Coeficientes do modelo - TIPVG_TTC

Preditor	Estimativas	Intervalo de Confiança a 95%		Erro-padrão	Z	p	Rácio das Chances	Intervalo de Confiança a 95%	
		Lim. Inferior	Superior					Lim. Inferior	Superior
Intercepto	−0.0113	−0.0446	0.0221	0.0170	−0.661	0.508	0.989	0.956	1.022
CG_TC	0.3477	0.2963	0.3992	0.0263	13.245	< .001	1.416	1.345	1.491
CG_TI	−0.5680	−0.6259	−0.5102	0.0295	−19.244	< .001	0.567	0.535	0.600
Pontuação Compoenente 1	0.5302	0.4936	0.5668	0.0187	28.390	< .001	1.699	1.638	1.763
Pontuação Compoenente 2	0.0806	0.0458	0.1154	0.0178	4.542	< .001	1.084	1.047	1.122
Pontuação Compoenente 3	−0.0977	−0.1322	−0.0633	0.0176	−5.553	< .001	0.907	0.876	0.939
Pontuação Compoenente 4	0.1224	0.0874	0.1573	0.0178	6.864	< .001	1.130	1.091	1.170

Nota. As estimativas representam o Log das Chances de "TIPVG_TTC = TI" vs. "TIPVG_TTC = TC"

Verificação de Pressupostos

Segundo Andy Field pág.914 - Valores de tolerância inferiores a 0,1 e VIF maiores do que 10 são preocupantes. Observando-se a tale a seguir, todas as variáveis são aprovadas no pressuposto de não multicolinearidade.

Estatísticas de Colinearidade

	VIF	Tolerância
CG_TC	2.35	0.425
CG_TI	2.54	0.394
Pontuação Compoenente 1	1.07	0.935
Pontuação Compoenente 2	1.12	0.889
Pontuação Compoenente 3	1.05	0.955
Pontuação Compoenente 4	1.09	0.917

[4]

Figura 13: intervalos de confiança para modelo de Regressão Logística sem ACP. Fonte: os autores.

CONCLUSÕES

Foi possível verificar que o uso conjunto das técnicas de Análise de Componentes Principais e Regressão Logit Binomial resultou em um modelo de previsão de divisão modal mais robusto. Principalmente no que se refere a:

1. Eliminação do problema de multicolinearidade das variáveis preditivas.
2. Melhora da qualidade preditiva a partir da redução da amplitude média dos intervalos de confiança dos coeficientes do modelo.
3. Incorporação de um conjunto de variáveis explicativas maior do que em um modelo sem a ACP.

REFERÊNCIAS BIBLIOGRÁFICAS

Artes, Rinaldo e Barroso, Lúcia. Métodos multivariados de análise estatística. São Paulo: Editora Edgard Blucher, 2023.

Damásio, Bruno Figueiredo. Uso da análise fatorial exploratória em Psicologia. Avaliação Psicológica, vol.11, no. 2 Itatiba. Abr./jun. 2012.
versão impressa ISSN 1677-0471*versão On-line* ISSN 2175-3431
http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712012000200007

Fávero, L.P.; Belfiore, P.; Silva, F.L.; Chan, B.L. Análise de Dados (modelagem multivariada para tomada de decisões). Rio de Janeiro: Elsevier - Edições Campus, 2009.

Fávero, L.P.; Belfiori, P. Manual de Análise de Dados. Estatística e modelagem multivariada com Excel, SPSS e Stata. Elsevier: Rio de Janeiro. 2017.

Fernandes, A. A. T. ; Filho, D. B. F.; Rocha, E. C.; Nascimento, W. S. Leia este artigo se você quiser aprender regressão logística. Revisa de Sociologia e Ciência Política, v.28, n. 74, e006, 2020.

Field, Andy. Descobrindo a estatística usando o SPSS. Artmed/Bookman: Porto Alegre, 2009.

Hosmer, David W., e Stanley Lemeschow. 2000. *Applied Logistic Regression*. 2 ed. New York: Wiley.

Kleinbaum, D.; Klein, M. Logistic Regression (a self-learning text) 3rd. edition. Springer: New York, Dordrecht, Heidelberg, London. Springer, 2010.

Lattin, J.; Carroll, J.D.; Green, P.E. Análise de dados multivariados. São Paulo: Cengage-Learning, 2011.

Lindner, Anabele; Pitombo, Cira S. Modelo logit binomial com componentes principais para estimação de preferência por modo de transporte motorizado. Journal of Transport Literature, 10(3), 5-9, Jul, 2016. IPTS (International Transport Planning Society).

Loesch, C.; Hoeltgebaum, M. Métodos Estatísticos Multivariados. Editora Saraiva: São Paulo, 2012.

Lopes, Andre. Medidas de performance: modelos de classificação. Publicado em 09/01/2023. https://brains.dev/2023/medidas-de-performance-modelos-de-classificacao/ acesso em 05/07/2024.

Maroco, João. Análise Estatística com utilização do SPSS, 3ª. edição. Edições Sílabo: Lisboa, 2007.

Miranda, Flávia Eloi. Utilização de Regressão Logística na Análise da Percepção da Comunidade da UFOP acerca da pandemia de Covid-19. Monografia de graduação. Depto de Estatística do Instituto de Ciências Exatas e Biológicas da UFOP. Orientador: Eduardo Bearzoti. Ouro Preto (MG), 2023.

Psicometria on line: Regressão logística – Pseudo-R^2. (13/03/2023) https://www.blog.psicometriaonline.com.br/regressao-logistica-pseudo-r%C2%B2/#:~:text=O%20Pseudo%20R%C2%B2%20%C3%A9%20uma,que%20%C3%A9

%20explicada%20pelo%20modelo.&text=Detalhando%20o%20Pseudo%20R%C2%B2%3A,m%C3%A1xima%20para%20o%20modelo%20completo. Acesso em 05/07/2024.

Reis, Elizabeth. Estatística Multivariada aplicada, 2ª. edição. Edições Sílabo: Lisboa, 2001.

Smolski, Felipe Micail da Silva; Battisti, Iara Denise Endruweit (2019). UFFS (Cerro Largo – extensão). Análise Fatorial com o R (item 7.3.2): <https://smolski.github.io/livroavancado/analisf.html> acesso em 05/07/2024.

Veras, Felipe. Abrindo a caixa preta: Análise de Componentes Principais. https://medium.com/@felipeverasaraujo/abrindo-a-caixa-preta-pca-an%C3%A1lise-de-componentes-principais-d5d400781dfe Medium. 18/07/2021. Acesso em 05/07/2024.