Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Matemática

Trabalho de Conclusão do Curso 2

Análise de componentes principais e ajustes de funções aos dados de COVID-19

Aluno: *Lucas Bastos Ricci Justino*

Curso: Bacharelado em Matemática

Orientador: *José Antonio Salvador*

Disciplina: Trabalho de conclusão de curso 2

Departamento de Matemática - DM - UFSCar

Periodo 2023/2

Trabalho de Conclusão do Curso 2

Análise de componentes principais e ajustes de funções aos dados de COVID-19

Aluno: *Lucas Bastos Ricci Justino*

Curso: Bacharelado em Matemática

Orientador: *José Antonio Salvador*

Disciplina: Trabalho de conclusão de curso 2

Departamento de Matemática - DM - UFSCar

Período 2023/2

Orientador: José Antonio Salvador	Orientando: Lucas Bastos Ricci Justino

Agradecimentos

A minha família por compreender minha ausência para que eu pudesse estar estudando em uma faculdade longe de casa e estar escrevendo este trabalho.

A meu orientador por ter me acolhido e me ajudado a desenvolver este trabalho.

Resumo

Neste Trabalho definimos o que são componentes principais de dados multivariados tanto quanto sua importância, suas propriedades e suas aplicações. Em seguida analisamos a possibilidade de aplicarmos o PCA em dados de Covid-19 obtidos através do SEADE (Sistema Estadual de Análise de Dados Estatísticos) e também exploramos com detalhes alguns ajustes de funções aos dados a fim de fazermos algumas observações pertinentes e apresentamos o modelo SIR(Suscetíveis, Infectados e Recuperados) para a exploração de doenças. Para melhor compreensão do leitor, usamos alguns conceitos básicos no primeiro capítulo, o qual são úteis durante o resto do trabalho.

Palavras-chave: Componentes principais. Covid-19. análise multivariada.

Abstract

In this work it is defined what are principal components of multivariate data as well as their importance, their properties and their applications. We will then analyze the possibility of applying the PCA to Covid-19 data obtained through SEADE (Sistema Estadual de Análise de Dados Estatísticos) we also explored in detail some function adjustments to the data in order to make some pertinent observations and presented the SIR (Susceptible, Infected and Recovered) model for exploring diseases. For the reader's better understanding, we use some basic concepts in the first chapter, which are useful throughout the rest of the work.

Key words: Principal component. Covid-19. Multivariate Analysis.

Sumário

Lista de Figuras

Prefácio

Em geral, para que serve o PCA? O PCA (Principal conponent analisis) ou APC (Análise de componentes principais) é uma técnica em análise de dados multivariados cujo objetivo é reduzir a dimensionalidade dos dados enquanto se preserva a maioria das informações, como a posição relativa de cada dado, variância e covariância de cada variável.

Ao diminuirmos a dimensionalidade dos dados também conseguimos diminuir sua complexidade computacional e aumentar consequentemente a velocidade de algorítimos que analisam tais dados. Além disso, pela natureza da construção das componentes principais, conseguimos retirar informações redundantes implicitadas de variáveis correlacionadas.

O PCA é melhor aplicado em bases cujos dados tenham uma alta quantidade de variáveis e de variáveis correlacionadas. Procuramos obter uma maior redução de dimensionalidade possível sem que muitas informações sejam perdidas. Buscamos neste trabalho estabelecer como é calculado o PCA e possíveis usos dele em dados de Covid-19.

Além disso, com o ajuste de funções e aos dados e alguns modelos matemáticos usando sistemas de EDO (Equações Diferenciais Ordinárias) podemos ter uma noção de como são feitos os gráficos e estatísticas sobre pandemias.

Capítulo 1

Conceitos Estatísticos Preliminares

Para utilizarmos o PCA é necessário o conhecimento prévio de alguns conceitos básicos estatísticos e matemáticos, os quais enunciaremos a seguir sem muito aprofundamento, pois não são o foco deste trabalho.

Definição 1.1. (Matriz de dados) Uma *matriz de dados* X é uma matriz (n × p) de tal forma que cada objeto analisado em todas suas variáveis são representados por uma linha n_i da matriz com i=1,2,...,n e cada uma das variáveis analisadas em cada objeto é uma coluna p_j com j=1,2,...,p. Desta forma a j-ésima variável do i-ésimo objeto será x_{ij} como vemos a seguir:

$$\mathrm{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Definição 1.2. (**Matriz transposta**) A transposta de uma matriz X denotada por X' é a matriz cujas linhas foram trocadas com as colunas

$$\mathrm{X}' = \begin{pmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1j} & \cdots & x_{ij} & \cdots & x_{nj} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{ip} & \cdots & x_{np} \end{pmatrix}$$

Exemplo 1.3. Seja A uma matriz de dimensões (2×3)

$$\mathrm{A} = \begin{pmatrix} 2 & 5 & 4 \\ 8 & 3 & 9 \end{pmatrix}, \text{ então } \mathrm{A}' = \begin{pmatrix} 2 & 8 \\ 5 & 3 \\ 4 & 9 \end{pmatrix}$$

Definição 1.4. (**Experimento aleatório**) Um evento que repetido inúmeras vezes com o mesmo processo gera resultados imprevisíveis é denominado um *experimento aleatório* (definiremos bem o que é um evento a seguir, mas inicialmente considere como conhecimento primitivo).

Denotaremos um experimento aleatório por E.

Exemplo 1.5. Jogar um dado e observar o número da face de cima é um experimento aleatório. Nos experimentos não se tem um resultado definido, mas ainda é possível saber todas as possibilidades, o que nos leva a próxima definição.

Definição 1.6. (**Espaço amostral**) Para cada experimento aleatório E é chamado de *espaço amostral*, denotado por S, o conjunto de todos os possíveis resultados de E.

Exemplo 1.7. Seja repetido o mesmo experimento anterior de se jogar o dado e observar o número de cima dele, então o espaço amostral é dado pelo conjunto de valores possiveis:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Definição 1.8. (**Evento**) Um conjunto de resultados do experimento, ou seja, um subconjunto de (S), é denominado um *evento*.

Exemplo 1.9. Novamente pensando no experimento de observar o resultado do lançamanto do dado, suponha o evento dado por A: tirar mais de 3 no dado, logo teremos o subconjunto dos valores possiveis

$$A = \{4, 5, 6\}$$

Definição 1.10. (**Probabilidade**) *Probabilidade* (geralmente entendida como a chance de algo ocorrer) é uma função P que tem domínio no espaço amostral E, que associe a cada evento A $\subseteq$ S a um número real entre 0 e 1 com P(S) = 1, ou seja

$$P : S \to [0, 1]$$

Exemplo 1.11. Utilizando o exemplo anterior de um dado qualquer temos que sendo $A = \{4, 5, 6\}$ temos que a probabilidade de ocorrer cada um de seus elementos é igual a $\frac{1}{6}$. Pois cada evento simples (tirar número específico no lançamento do dado) tem mesma probabilidade, assim conseguimos encontrar a probabilidade de A através da soma:

$$P(A) = P(4) + P(5) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Definição 1.12. (**Variável aleatória**) Seja E um experimento aleatório e S o espaço amostral de E, é denominada *variável aleatória* uma função X que relaciona cada elemento s $\in$ S a um número real X(s).

Uma variável aleatória X é dita *discreta* se o seu contradomínio for um conjunto finito (quando o conjunto é vazio ou quando existe n natural tal que exista uma bijeção entre, tal conjunto e $\{1, ..., n\}$) ou enumerável (quando existe uma bijeção entre tal conjunto e o conjunto dos números naturais) caso ele seja um intervalo ou a união de intervalos então ela é denominada *contínua*.

Exemplo 1.13. Seja o experimento E de se jogar uma moeda três vezes e observar se o resultado será cara ou coroa, com isso teremos para cada três lançamentos as possibilidades indicadas como elementos de S,

$$S = \{(cara, cara, cara), (cara, cara, coroa), (cara, coroa, cara), (cara, coroa, coroa),$$
$$(coroa, cara, cara), (coroa, cara, coroa), (coroa, coroa, cara), (coroa, coroa, coroa)\}$$

Podemos ter como variável aleatória a função X: número de vezes que o resultado foi cara. assim fica claro que para cada s ∈ S teremos os números reais 0, 1, 2, 3 associados a eles de tal forma que

$$X((cara, cara, coroa)) = 2; X((coroa, coroa, coroa)) = 0; X((cara, coroa, coroa)) = 1$$

Reciprocamente para as outras possibilidades de E.

Como o contradomínio de X é um conjunto finito de naturais, então X é uma variável aleatória discreta.

Definição 1.14. (**Média aritmética**) *Média aritmética* denotada por $\overline{x}$ é a soma de todos os n elementos de um conjunto dividido pela quantidade de elementos que esse conjunto possui.

$$\overline{x} = \frac{\sum_{i=1}^{n} a_i}{n}$$

Exemplo 1.15. Suponha que nosso conjunto seja composto pela quantidade de novos casos de Covid diários registrados no Estado de São Paulo entre os dias 21 de julho até 3 de agosto de 2020, conforme a Figura 1.1 a seguir.

Logo a média aritmética $\overline{x}$ será dada por

$$\overline{x} = \frac{6608 + 6745 + ... + 6860 + 5970}{14} = 5130$$

Podemos observar que neste ultimo exemplo aplicado existe uma notável diferença entre a quantidade de casos nos fins de semana e em dias de semana, isso se da por fatores humanos como fato do registro de casos ser pouco acurada já que parte dos registros do fim de semana só volta a ser feita na segunda-feira da outra semana. Desse modo ao analisarmos dados referentes a infectados, e óbitos é conveniente usarmos sempre médias semanais moveis ao invés dos valores diários.

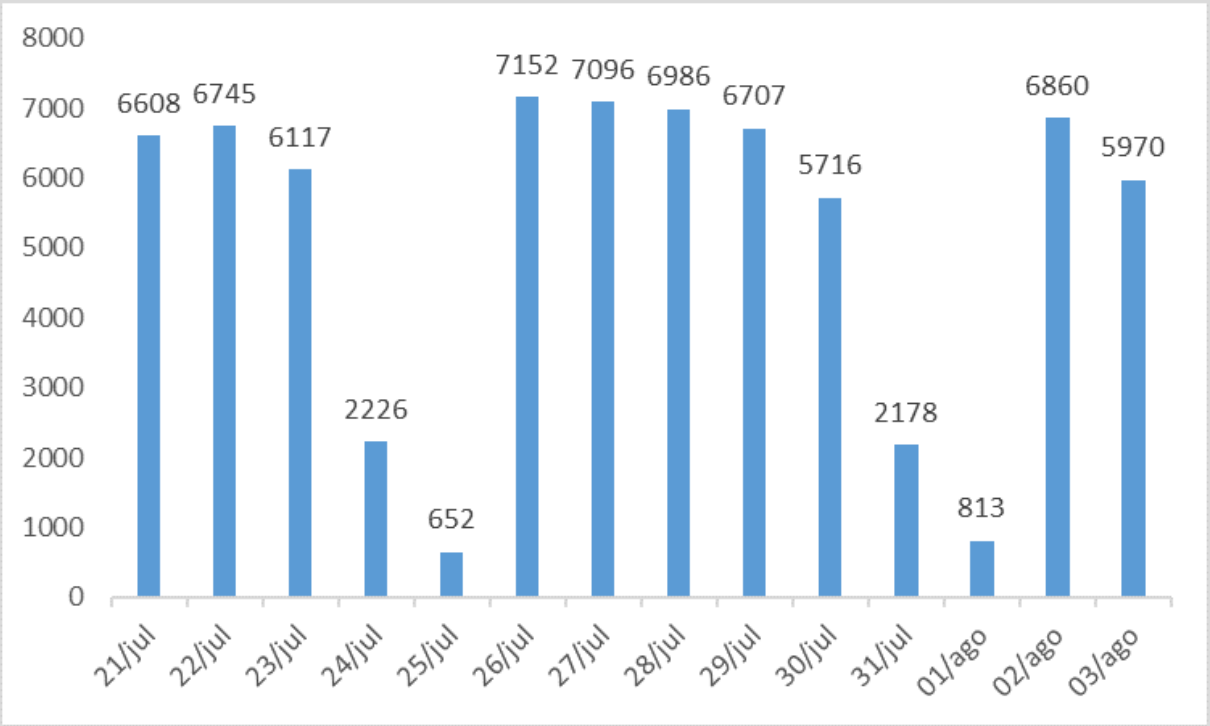


Figura 1.1: Casos confirmados de Covid-19. 21/jul - 03 ago 2022 SEADE

Definição 1.16. (**Média ou esperança**) É definido como *esperança matemática* de uma Variável aleatória discreta E[X] (também denotada por μ_X) pela equação:

$$E[X] = \sum_i x_i \ P(x_i)$$

Por esta definição observamos que quando todos os eventos tiverem a mesma probabilidade o valor esperado será a própria média aritmética $\overline{x}$, ou seja:

$$E[X] = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1+x_2+...+x_n}{n} = \overline{x}$$

com n sendo o número de eventos do espaço S.

Definição 1.17. (**Variância**) É definido a *variância* de uma variável aleatória X denotada por $Var[X]$ (ou denotada por σ^2) como:

$$\sigma^2 = Var[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

a variância é diretamente relacionada com o desvio-padrão como veremos a seguir

Definição 1.18. (**Desvio-padrão**) É definido por *desvio-padrão* denotado por σ a raiz quadrada da variância, ou seja

$$\sigma = \sqrt{Var[X]}$$

Definição 1.19. (**Covariância**) A Covariância entre duas variáveis aleatórias X e Y é dada por:

$$\text{Cov}(X, Y) = E[X - E(X)][Y - E(Y)]$$

Estendendo os conceitos de média, variância e covariância de elementos com apenas uma variável para elementos com várias variáveis temos as seguintes definições:

Definição 1.20. (**Média multivariada**) A média amostral da i-ésima variável é dada por:

$$\overline{x}_i = \frac{1}{n}\sum_{r=1}^{n} x_{ri}$$

Definição 1.21. (**Variância multivariada**) A variância amostral da i-ésima variável é dada por:

$$s_{ii} = \frac{1}{n}\sum_{r=1}^{n}(x_{ri} - \overline{x}_i)^2 = s_i^2, \text{ com } i = 1, \ldots, p$$

Definição 1.22. (**Covariância multivariada**) Dadas as duas ultimas definições podemos definir a covariância amostral entre a i-ésima e a j-ésima variável, dada por:

$$s_{ij} = \frac{1}{n}\sum_{r=1}^{n}(x_{ri} - \overline{x}_i)(x_{rj} - \overline{x}_j)$$

Definição 1.23. (**Vetor coluna**) Definimos como vetor coluna $\mathbf{x}$ a matriz $(n \times 1)$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_p \end{pmatrix}$$

Podemos visualizar tal matriz como um vetor $\mathbf{x} = (x_1, \cdots, x_j, \cdots, x_p)\prime$ cujas suas coordenadas no espaço p-dimensional são representadas por cada uma das variáveis analisadas no elemento $\mathbf{x}$

Exemplo 1.24. (**Vetor das médias**) Definimos agora o vetor das médias, de forma que cada um de seus elementos seja a média entre todos os objetos de cada j-ésima variável

$$\overline{\mathbf{x}} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_j \\ \vdots \\ \overline{x}_p \end{pmatrix}$$

Definição 1.25. (**Matriz de covariância**) A matriz quadrada S (p × p) que tem como cada elemento a covariância entre as i-ésimas e as j-ésimas variáveis é chamada de matriz de covariância

$$\mathbf{S} = \begin{pmatrix} s_{11} & \cdots & s_{1j} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{i1} & \cdots & s_{ij} & \cdots & s_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pj} & \cdots & s_{pp} \end{pmatrix}$$

Combinações lineares são uma importante ferramenta de análise multivariada, criando combinações lineares com variáveis dos elementos analisados é possível tornar a interpretação dos dados mais clara.

Definição 1.26. (**Combinação linear**) Dado o elemento x_r a ser analisado, uma combinação linear y_r de suas variáveis é da forma

$$y_r = a_1 x_{r1} + \cdots + a_1 x_{r1};\ r = 1, \cdots, n;\ a_r \in \mathbb{R}$$

Definição 1.27. (**Vetor unitário**) Um vetor unitário $\hat{v}$ de um vetor $\vec{v}$ é um vetor que tem mesma direção de $\vec{v}$ e que tenha comprimento igual a 1.

$$\hat{v} = \frac{\vec{v}}{|\vec{v}|}$$

Onde |v| é a norma de v.

Definição 1.28. (**Autovalor e Autovetor**)

Dada uma matriz quadrada X, um número escalar λ é um **autovalor** de X se existir um vetor não nulo v tal que

$$Xv = \lambda v$$

O vetor v é chamado de **autovetor** correspondente a λ. Dado uma matriz X a equação característica é dada por

$$\det(X - \lambda I) = 0$$

onde I é a matriz identidade.

Para cada autovalor λ_i, para encontrar os autovetores podemos resolver o sistema de equações:

$$(X - \lambda_i I)\mathbf{v}_i = \mathbf{0}$$

onde $\mathbf{v}_i$ é o autovetor correspondente a λ_i.

Exemplo 1.29. Considere a matriz

$$X = \begin{pmatrix} 3 & 1 & 0 \\ -1 & 2 & -1 \\ 1 & -1 & 4 \end{pmatrix}$$

A equação característica é

$$\det(X - \lambda I) = \det\left(\begin{pmatrix} 3-\lambda & 1 & 0 \\ -1 & 2-\lambda & -1 \\ 1 & -1 & 4-\lambda \end{pmatrix}\right) = (\lambda - 3)(\lambda - 2)(\lambda - 4) = 0$$

Os autovalores são $\lambda_1 = 3$, $\lambda_2 = 2$, e $\lambda_3 = 4$.

Encontrando os Autovetores

Para $\lambda_1 = 3$, temos

$$\begin{pmatrix} 0 & 1 & 0 \\ -1 & -1 & -1 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Para $\lambda_2 = 2$, temos

$$\begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & -1 \\ 1 & -1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$$

Para $\lambda_3 = 4$, temos

$$\begin{pmatrix} -1 & 1 & 0 \\ -1 & -2 & -1 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$v_3 = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$$

Existem muitos outros conceitos estatísticos e algébricos importantes que podemos explorar, porem os expostos neste capitulo com alguns outros que aparecerão são suficientes para trabalharmos na construção do PCA.

Capítulo 2

Conceitos Sobre o PCA

No processo da análise de componentes principais, um conjunto de obervações (variáveis) possivelmente relacionadas é transformado em um conjunto linear não correlacionado (linearmente independente). Este conjunto de variáveis não relacionadas são as chamadas componentes principais. O número de componentes principais é sempre menor ou igual o número de variáveis.

Suponhamos um vetor aleatória x da forma

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_n \end{pmatrix}$$

E suponhamos as seguintes combinações lineares:

$$\begin{cases} P_1 = \xi_{11}x_1 + \xi_{12}x_2 + ... + \xi_{1n}x_n \\ P_2 = \xi_{21}x_1 + \xi_{22}x_2 + ... + \xi_{2n}x_n \\ \vdots \\ P_n = \xi_{n1}x_1 + \xi_{n2}x_2 + ... + \xi_{nn}x_n \end{cases}$$

Cada P_i pode ser considerado como uma regressão linear, ou seja, P_i pode ser deduzida a partir de $x_1, x_2...x_n$ e $\xi_{11}, \xi_{12}, ..., \xi_{1n}$, como os coeficientes da regressão. Como P_i é combinação de observações aleatórias, então P_i é também aleatório, de forma que podemos calcular suas variâncias e covariâncias.

Entre todas as combinações lineares consideramos a primeira componente principal (PC1) a que tiver a maior variância possível, ou seja, os coeficientes $\xi_{11}, \xi_{12}, ..., \xi_{1n}$, são definidos de maneira a maximizar a variância, com a restrição de que a soma dos quadrados desses coeficientes seja igual a um, assim existira uma única solução possível

Em seguida a segunda componente principal (PC2)

Entre todas as combinações lineares, PC2 é a combinação linear das variáveis que explica o máximo possível da variação restante nos dados, tendo à restrição de que a PC1 e PC2 não tenham correlação. E novamente a soma dos quadrados seja igual a 1. Escolhendo adequadamente os coeficientes $\xi_{21}, \xi_{22}, \ldots, \xi_{2\rho}$ que maximizam a variância deste novo componente:

$$\text{var}(P_2) = \sum_{k=1}^{n}\sum_{l=1}^{n} \xi_{2k}\xi_{2l}s_{kl} = \xi_2'\mathbf{S}\xi_2,$$

sujeita à restrição de que

$$\xi_2'\xi_2 = \sum_{j=1}^{n} \xi_{2j}^2 = 1.$$

junto com a restrição adicional de que a correlação entre esses dois componentes é 0:

$$\text{cov}(P_1, P_2) = \sum_{k=1}^{n}\sum_{l=1}^{n} \xi_{1k}\xi_{2l}s_{kl} = \xi_1'\mathbf{S}\xi_2 = 0.$$

Todas as PCs restantes têm a mesma propriedade específica - todas as PCs são combinações lineares que explicam o máximo possível da variação restante no conjunto de dados, e todas as PCs não estão correlacionadas entre si. Portanto, a PCi pode ser derivada da seguinte forma: Escolhemos adequadamente os coeficientes $\xi_{i1}, \xi_{i2}, \ldots, \xi_{i\rho}$ que maximizam

$$\text{var}(P_i) = \sum_{k=1}^{n}\sum_{l=1}^{n} \xi_{ik}\xi_{ij}s_{kl} = \xi_i'\mathbf{S}\xi_i,$$

sujeita à restrição de que a soma dos coeficientes ao quadrado é 1, juntamente com a restrição adicional de que a correlação entre esta nova PC e todas as PCs anteriores é 0:

$$\xi_i'\xi_i = \sum_{j=1}^{n} \xi_{ij}^2 = 1,$$

$$\text{cov}(P_1, P_i) = \sum_{k=1}^{n}\sum_{l=1}^{n} \xi_{1k}\xi_{ij}s_{kl} = \xi_1'\mathbf{S}\xi_i = 0,$$

$$\text{cov}(P_2, P_i) = \sum_{k=1}^{n}\sum_{l=1}^{n} \xi_{2k}\xi_{ij}s_{kl} = \xi_2'\mathbf{S}\xi_i = 0,$$

$$\vdots$$

$$\text{cov}(P_{i-1}, P_i) = \sum_{k=1}^{n}\sum_{l=1}^{n} \xi_{i-1,k}\xi_{ij}s_{kl} = \xi_{i-1}'\mathbf{S}\xi_i = 0.$$

Consequentemente, todas as PCs não são correlacionadas entre si. Portanto, PC1 explica a maior variância possível, ou seja, representa o máximo da variabilidade nos dados originais.

Capítulo 3

Aplicação do PCA

Imaginemos a situação de um professor de escola que deseja observar a variedade das aptidões dos alunos, dadas as seguintes notas das matérias, com um trabalho multidisciplinar especial.

Aluno/Matéria	Português	Matemática	Ciências	História	Geografia	Trabalho E.	Média final
Maurício	8	8	7	8	6,5	9,5	7,83
Paula	8	9	7	10	6	10	8,33
Augusto	7	7	6	8	9	10	7,83
Vitoria	10	7,5	10	7,5	6,5	9	8,42
Tatiana	8	9	8,5	6	7	9,5	8,00
Tiago	8	7	8,5	6	8,5	9	7,83

Tabela 3.1: Quadro das notas dos alunos

De quais maneiras podemos observar tal variedade, uma das maneiras seria a observação da própria média final, mas será que é possível obtermos alguma informação desta maneira? Outra forma seria para cada dois alunos calcularmos quão distantes eles estão considerando cada nota como uma de variável de um ponto no espaço, mas novamente essa analise seria problemática, pois com a quantidade de relações seria inviável analisarmos todas elas simultaneamente. Esse é o tipo de questão o qual a PCA busca resolver.

Estamos procurando uma transformação de mudança de base que leve os elementos analisados para uma base em que a variância relativa à primeira componente seja a maior possível que a componente seguinte, seja ortogonal a primeira e novamente com a maior variância possível e assim por diante, deste modo teremos no total o mesmo número p de variáveis que tínhamos inicialmente, porem a maioria da variação dos dados estará presente nas primeiras componentes, desse modo poderemos descartar quantas componentes forem convenientes sem grande perda de informação significativa para o problema.

Para aplicarmos o PCA pode haver a possibilidade dos dados serem padronizados ou normalizados, em seguida representemos cada um dos n elementos como vetores de p variáveis no R^p.

Vamos pensar nos dados como uma matriz n×p, neste caso 6×6 da forma

$$\mathrm{X} = \begin{pmatrix} 8 & 8 & 7 & 8 & 6,5 & 9,5 \\ 8 & 9 & 7 & 10 & 6 & 10 \\ 7 & 7 & 6 & 8 & 9 & 10 \\ 10 & 7,5 & 10 & 7,5 & 6,5 & 9 \\ 8 & 9 & 8,5 & 6 & 7 & 9,5 \\ 8 & 7 & 8,5 & 6 & 8,5 & 9 \end{pmatrix}$$

Primeiro devemos calcular as médias e desvios padrão de cada característica, neste caso de cada matéria em questão.

$$\overline{\mathbf{x}} = \begin{pmatrix} 8,17 \\ 7,92 \\ 7,83 \\ 7,58 \\ 7,25 \\ 9,5 \end{pmatrix} \; ; \; \sigma = \begin{pmatrix} 0,98 \\ 0,92 \\ 1,44 \\ 1,50 \\ 1,21 \\ 0,45 \end{pmatrix}$$

Em seguida padronizamos as variáveis, ou seja, deixaremos cada uma das variáveis tendo média = 0 e desvio padrão = 1, para fazer isto basta subtrair de cada elemento sua respectiva média e dividir pelo seu respectivo desvio padrão.

$$p_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}$$

Assim a matriz de dados padronizados fica

$$X_p = \begin{pmatrix} -0,17 & 0,09 & -0,58 & 0,28 & -0,62 & 0 \\ -0,17 & 1,18 & -0,58 & 1,61 & -1,03 & 1,12 \\ -1,19 & -1,00 & -1,28 & 0,28 & 1,44 & 1,12 \\ 1,86 & -0,45 & 1,51 & -0,06 & -0,62 & -1.12 \\ -0,17 & 1,18 & 0,46 & -1,06 & -0,21 & 0 \\ -0,17 & -1,00 & 0,46 & -1,06 & 1,03 & -1,12 \end{pmatrix}$$

Agora faremos a matriz de covariância dos dados a partir dos dados obtidos

$$\mathbf{S} = \begin{pmatrix} 0,83 & 0,02 & 0,73 & -0,07 & -0,45 & -0,57 \\ 0,02 & 0,83 & -0,01 & 0,25 & -0,62 & 0,3 \\ 0,73 & -0,01 & 0,83 & -0,42 & -0,24 & -0,71 \\ -0,07 & 0,25 & -0,42 & 0,83 & -0,38 & 0,56 \\ -0,45 & -0,62 & -0,24 & -0,38 & 0,83 & 0 \\ -0,57 & 0,3 & -0,71 & 0,56 & 0 & 0,83 \end{pmatrix}$$

Iremos agora calcular os autovalores de **S** utilizando o polinômio característico dados por

$$\mathrm{p}(\lambda) = \det\,(\mathbf{S} - \mathbf{I}\lambda)$$

$$(\mathbf{S} - \mathbf{I}\lambda) = \begin{pmatrix} 0,83-\lambda & 0,02 & 0,73 & -0,07 & -0,45 & -0,57 \\ 0,02 & 0,83-\lambda & -0,01 & 0,25 & -0,62 & 0,3 \\ 0,73 & -0,01 & 0,83-\lambda & -0,42 & -0,24 & -0,71 \\ -0,07 & 0,25 & -0,42 & 0,83-\lambda & -0,38 & 0,56 \\ -0,45 & -0,62 & -0,24 & -0,38 & 0,83-\lambda & 0 \\ -0,57 & 0,3 & -0,71 & 0,56 & 0 & 0,83-\lambda \end{pmatrix}$$

Tendo o polinômio característico conseguimos calcular os autovalores

$$\Lambda_S = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \end{pmatrix} = \begin{pmatrix} 2,47 \\ 1,82 \\ 0,58 \\ 0,10 \\ 0,03 \\ 0 \end{pmatrix}$$

e consequentemente os autovetores que de forma ordenada formam a matriz

$$V_S = \begin{pmatrix} -0,48 & -0,30 & -0,38 & -0,31 & -0,50 & 0,44 \\ 0,12 & -0,55 & 0,65 & -0,16 & -0,41 & -0,26 \\ -0,56 & -0,12 & 0,07 & -0,49 & 0,53 & -0,37 \\ 0,35 & -0,38 & -0,65 & -0,05 & -0,06 & -0,54 \\ 0,09 & 0,66 & 0 & -0,51 & -0,45 & -0,32 \\ 0,56 & -0,12 & 0,02 & -0,61 & 0,31 & 0,45 \end{pmatrix}$$

Em que cada coluna representa um autovetor y_j. Tais autovetores são nossos componentes principais. Em seguida, para sabermos as coordenadas dos nossos dados padronizados, multiplicaremos a matriz dos dados com a matriz dos autovetores que acabamos de calcular.

$$X_p \times V_S$$

$$\begin{pmatrix} -0,17 & 0,09 & -0,58 & 0,28 & -0,62 & 0 \\ -0,17 & 1,18 & -0,58 & 1,61 & -1,03 & 1,12 \\ -1,19 & -1,00 & -1,28 & 0,28 & 1,44 & 1,12 \\ 1,86 & -0,45 & 1,51 & -0,06 & -0,62 & -1.12 \\ -0,17 & 1,18 & 0,46 & -1,06 & -0,21 & 0 \\ -0,17 & -1,00 & 0,46 & -1,06 & 1,03 & -1,12 \end{pmatrix} \times$$

$$\begin{pmatrix} -0,48 & -0,30 & -0,38 & -0,31 & -0,50 & 0,44 \\ 0,12 & -0,55 & 0,65 & -0,16 & -0,41 & -0,26 \\ -0,56 & -0,12 & 0,07 & -0,49 & 0,53 & -0,37 \\ 0,35 & -0,38 & -0,65 & -0,05 & -0,06 & -0,54 \\ 0,09 & 0,66 & 0 & -0,51 & -0,45 & -0,32 \\ 0,56 & -0,12 & 0,02 & -0,61 & 0,31 & 0,45 \end{pmatrix} =$$

$$\begin{pmatrix} -0,46 & 0,44 & -0,1 & 0,62 & -0,16 & 0 \\ -1,65 & 1,96 & -0,25 & -0,09 & 0,21 & 0 \\ -2,01 & -1,76 & -0,46 & -0,29 & -0,12 & 0 \\ 2,49 & 0,75 & -0,87 & -0,24 & -0,09 & 0 \\ 0,42 & 0,38 & 1,56 & -0,21 & -0,08 & 0 \\ 1,21 & -1,76 & 0,12 & 0,2 & 0,25 & 0 \end{pmatrix}$$

Cada uma das componentes contem uma porcentagem da variância, tal porcentagem é calculada através da divisão de cada relativo autovalor pela soma de todos os autovalores,

$$\text{porcentagem da variância explicitada} = \frac{\lambda_i}{\sum_{k=1}^{n} \lambda_k}$$

assim temos

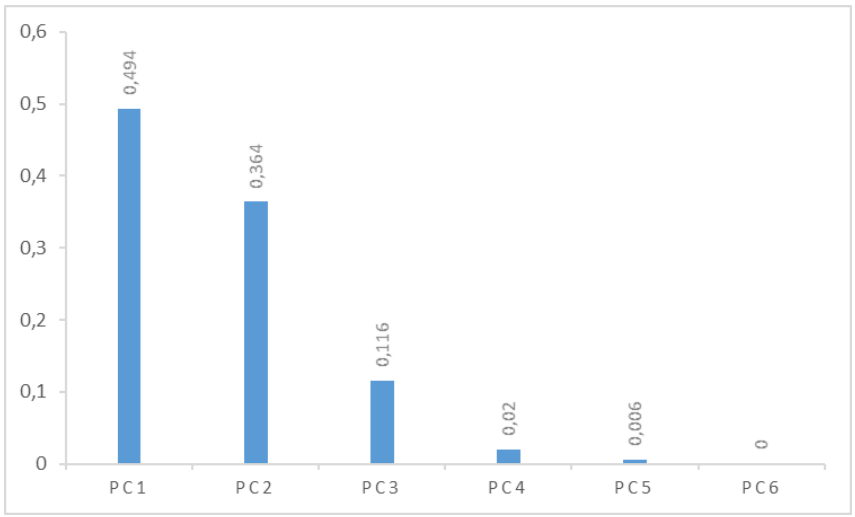


Figura 3.1: Representação da variância de cada componente).

Sendo cada elemento da amostra (neste caso cada aluno) representada por uma linha, como queremos reduzir a dimensionalidade dos dados para ser possível a visualização gráfica deles, irmos reduzir a matriz V_s criando uma matriz apenas com duas e em seguida com três variáveis, assim

$$X_p \times V_{S2}$$

$$\begin{pmatrix} -0,17 & 0,09 & -0,58 & 0,28 & -0,62 & 0 \\ -0,17 & 1,18 & -0,58 & 1,61 & -1,03 & 1,12 \\ -1,19 & -1,00 & -1,28 & 0,28 & 1,44 & 1,12 \\ 1,86 & -0,45 & 1,51 & -0,06 & -0,62 & -1.12 \\ -0,17 & 1,18 & 0,46 & -1,06 & -0,21 & 0 \\ -0,17 & -1,00 & 0,46 & -1,06 & 1,03 & -1,12 \end{pmatrix} \begin{pmatrix} -0,48 & -0,30 \\ 0,12 & -0,55 \\ -0,56 & -0,12 \\ 0,35 & -0,38 \\ 0,09 & 0,66 \\ 0,56 & -0,12 \end{pmatrix} =$$

$$\begin{pmatrix} -0,46 & 0,44 \\ -1,65 & 1,96 \\ -2,01 & -1,76 \\ 2,49 & 0,75 \\ 0,42 & 0,38 \\ 1,21 & -1,76 \end{pmatrix}$$

Com duas componentes representamos 85,8% da variância original

Analogamente para três componentes principais teremos

$$X_p \times V_{S3}$$

$$\begin{pmatrix} -0,17 & 0,09 & -0,58 & 0,28 & -0,62 & 0 \\ -0,17 & 1,18 & -0,58 & 1,61 & -1,03 & 1,12 \\ -1,19 & -1,00 & -1,28 & 0,28 & 1,44 & 1,12 \\ 1,86 & -0,45 & 1,51 & -0,06 & -0,62 & -1.12 \\ -0,17 & 1,18 & 0,46 & -1,06 & -0,21 & 0 \\ -0,17 & -1,00 & 0,46 & -1,06 & 1,03 & -1,12 \end{pmatrix} \times \begin{pmatrix} -0,48 & -0,30 & -0,38 \\ 0,12 & -0,55 & 0,65 \\ -0,56 & -0,12 & 0,07 \\ 0,35 & -0,38 & -0,65 \\ 0,09 & 0,66 & 0 \\ 0,56 & -0,12 & 0,02 \end{pmatrix} =$$

$$\begin{pmatrix} -0,46 & 0,44 & -0,1 \\ -1,65 & 1,96 & -0,25 \\ -2,01 & -1,76 & -0,46 \\ 2,49 & 0,75 & -0,87 \\ 0,42 & 0,38 & 1,56 \\ 1,21 & -1,76 & 0,12 \end{pmatrix}$$

representando agora 97,4% da variância.

Não existe um consenso de quantos amostras precisamos utilizar para que o PCA seja considerado viável, porem, em geral, ele é utilizado quando temos 5 vezes mais amostras que variáveis ou no mínimo 100 amostras para que ela seja justificada. Mesmo que os dados utilizados neste exemplo fujam de tal suposição, com eles conseguimos visualizar a forma de se calcular os componentes principais.

Plotando o gráfico para duas e três componentes:

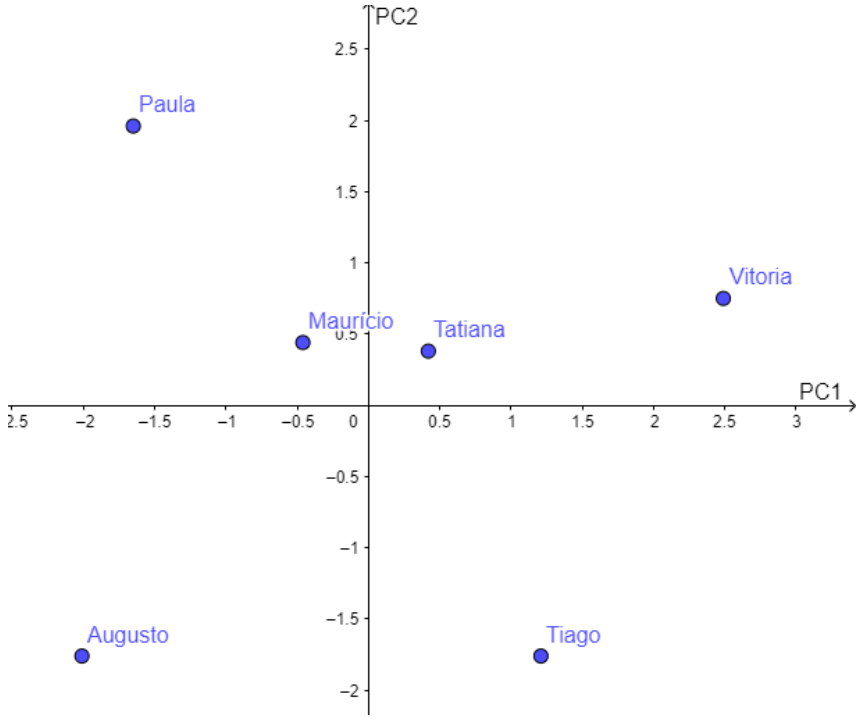


Figura 3.2: Distribuição com duas componentes (85,8% da variância representada).

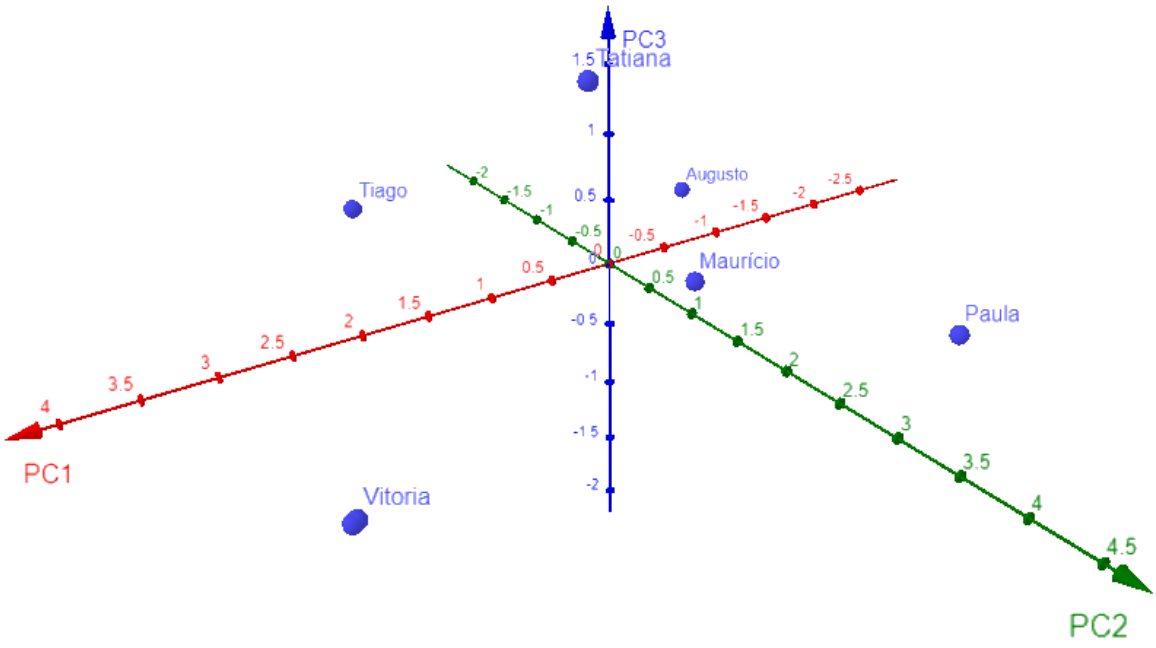


Figura 3.3: Distribuição com três componentes (97,4% da variância representada).

Capítulo 4

PCA Para COVID-19

Para aplicarmos o PCA em dados de Covid precisamos primeiramente encontrar uma fonte com uma base de dados confiável e apropriada para tal. Desta forma foi escolhido por utilizar dos dados do SEADE (Sistema Estadual de Análise de Dados Estatísticos) sendo a fundação vinculada a secretaria do governo que atua, produzindo, analisando e distribuindo dados estatísticos de caráter econômico, social, demográfico entre outros para todo o Brasil.

O SEADE disponibiliza de forma acessível os dados sobre Covid em seu site (seade.gov.br/coronavirus), onde estão explícitos dados e gráficos de casos confirmados, óbitos, variação diária, letalidade, entre outros, todos separados pelas suas respectivas cidades, estados e países. Juntamente existem links diretos para se obter mais informações sobre outras cidades, estados, países. Além disso, é disponível para download a base de dados coletada dia após dia desde o começo da pandemia até os dias atuais.

Nos repositórios de dados do SEADE encontramos cinco categorias de bancos de dados que iremos analisar, sendo estes: *Casos e óbitos por município e data*; *Leitos e internações*; *Hospitalizados por síndrome respiratória aguda*; *casos e óbitos por doenças pré-existentes, sexo e idade*; *Casos e óbitos por raça/cor e município*[5].

4.1 Casos e óbitos

Olhando para os dados de casos e óbitos temos como elemento analisado cada município, e como variáveis que podemos trabalhar o número de casos por habitantes, óbitos por habitantes, e letalidade. Outras variáveis úteis que podemos trabalhar são as médias móveis de casos e de óbitos, porem como cada município tem uma grande diferença de população é necessário trabalhar estas variáveis a fim de obter valores de magnitudes não tão diferentes. Outras variáveis como código de município e nome de DRS (Departamentos Regionais de Saúde) não tem relevância, pois estão atrelados a cada município.

4.2 Leitos e Internações

Neste banco de dados temos cada DRS como elemento analisado, e como variáveis que podemos trabalhar ultilizando a ocupação de leitos, leitos por pessoa, número de internações, número de internações em leitos de enfermaria.

4.3 SRAG

Neste Banco de dados temos os pacientes como elementos a observar e mais de 130 variáveis que vão de nomes de município até a quantidade de cada respectivos sintomas quando internados, apesar disso ela esta desatualizada com últimos registros datando junho de 2020, mesmo assim foram registrados mais de 27 mil hospitalizações, sendo assim um ótimo banco de dados para aplicarmos o PCA.

4.4 Doenças pré-existentes

Neste banco de dados temos como cada elemento um paciente e como variáveis que podemos utilizar temos: idade, sexo, confirmação de Covid-19, óbito e por fim cada um dos tipos de fatores de risco como, por exemplo, asma, diabetes, obesidade entre outros. Este é o Banco de dados com mais fatores binários (ou se tem o fator de risco ou não), sendo assim torna-se viável a separação dos pacientes por categorias que podem ser previstas através da sua distribuição ao aplicarmos a PCA.

4.5 Raça/cor e município

Esta base possui as mesmas variáveis que a da de doenças pré-existentes, porem com a adição da raça/cor e com a respectiva DRS, e por isso possui possibilidades de análise semelhante.

4.6 Aplicando o PCA

Optando pelo uso dos dados de SRAG podemos verificar, por exemplo, se dados um conjunto de sintoma dos pacientes podemos Inferir com uma probabilidade de certeza que tal paciente teve problemas respiratórios por causa específica de covid, desta forma iremos comparar a ocorrência de cada sintoma registrado e a confirmação da causa da Síndrome respiratória, ambas informações presentes nos dados de SRAG.

Foram registrados 106.651 pacientes nos dados nos quais foram utilizados para esta análise apenas aqueles os quais possuem todas as entradas essenciais preenchidas, (ou seja, tanto as entradas dos sintomas quanto a confirmação de covid preenchidas). Foram

observados 11 fatores, sendo estes: caso proveniente de uma síndrome gripal que evoluiu para SRAG; adquiriu a doença no hospital; paciente trabalha ou tem contato direto com aves ou suínos; Sinais e sintomas de febre; Sinais e sintomas de tosse; Sinais e sintomas de dor de garganta; Sinais e sintomas de dor de Dispneia; Sinais e sintomas de dor de desconforto respiratório; Sinais e sintomas de Saturação; Sinais e sintomas de Diarreia; Sinais e sintomas de vômito. Utilizando o sistema binário de 1 caso afirmativo e 0 caso negativo, e junto para cada caso a confirmação de se a doença foi ou não causada por Covid (como na figura a baixo).

	SURTO_SG	NOSOCOMIAL	AVE_SUINO	FEBRE	TOSSE	GARGANTA	DISPNEIA	DESC_RESP	SATURACAO	DIARREIA	VOMITO	CLASSI_FIN
1	0	0	0	1	1	0	1	1	0	0	1	Outro
2	0	0	0	1	1	0	1	1	0	0	0	Outro
3	0	0	0	0	1	0	1	1	1	1	1	Outro
4	0	0	0	0	1	0	1	1	1	0	0	Outro
5	0	0	0	1	1	1	1	1	0	1	0	Outro
6	0	0	0	1	1	1	1	1	0	0	0	Outro
7	0	0	0	1	1	0	1	1	1	0	0	Outro
8	0	0	0	0	1	0	1	1	1	0	0	Outro
9	0	0	0	0	1	0	1	1	1	0	0	Outro
10	0	0	0	0	0	0	1	1	1	0	0	Outro
11	0	0	0	1	1	0	1	1	0	0	0	Outro
12	0	0	0	1	1	0	1	1	1	0	1	Outro
13	0	0	0	1	1	0	1	1	0	0	0	Outro
14	0	0	0	1	1	0	1	1	1	0	1	Outro
15	0	0	0	1	1	0	1	1	1	0	0	Outro
16	0	0	0	1	1	0	1	1	1	0	0	Outro
17	0	0	0	1	1	0	1	1	1	0	0	Outro

Showing 1 to 5 of 106,651 entries, 12 total columns

Figura 4.1: Registro de dados selecionados - SRAG

Trabalhando com os Dados no Rstudio e construindo as 11 componentes principais(relativas as 11 variáveis) temos a matriz de correlação e a representação de cada componente principal

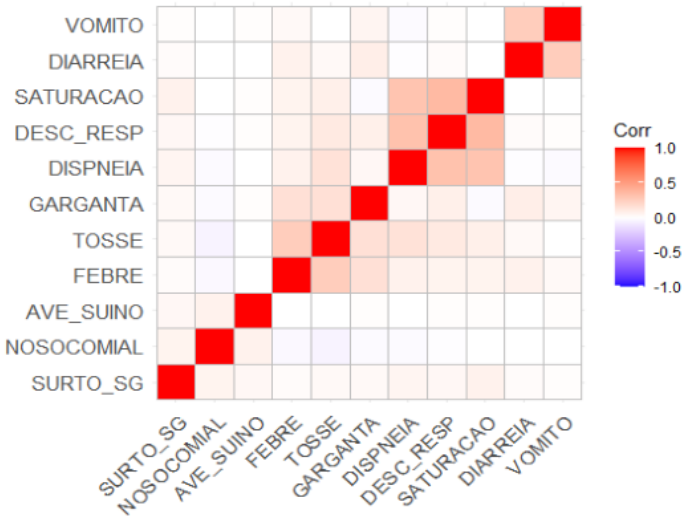


Figura 4.2: Matriz correlação - dados SRAG - R).

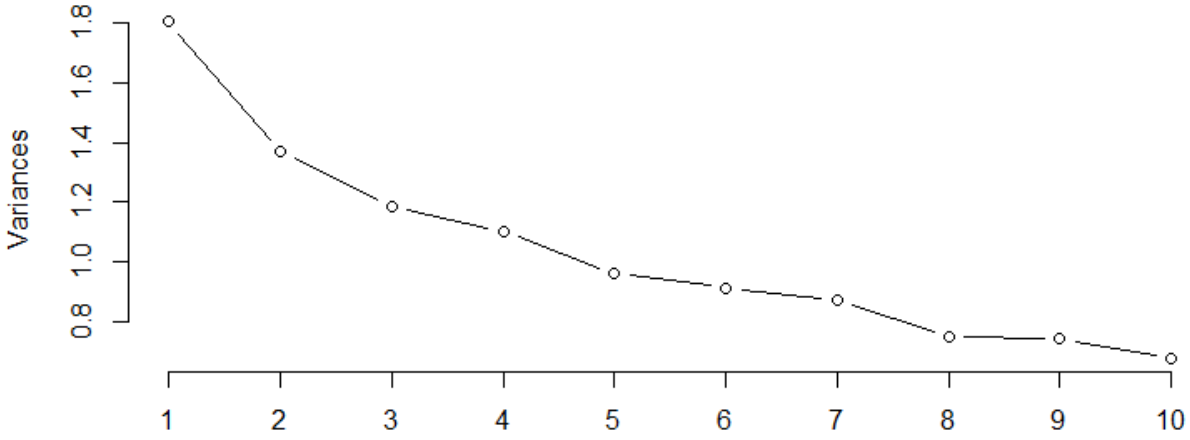


Figura 4.3: Representação da variância de cada componente - R

Observamos que as duas primeiras componentes representam cerca de apenas 29% da variância, também podemos observar a baixa correlação entre os sintomas na matriz de correlação.

Colocando os dados agora com base nas duas primeiras componentes principais podemos observar a aglomeração dos dados.

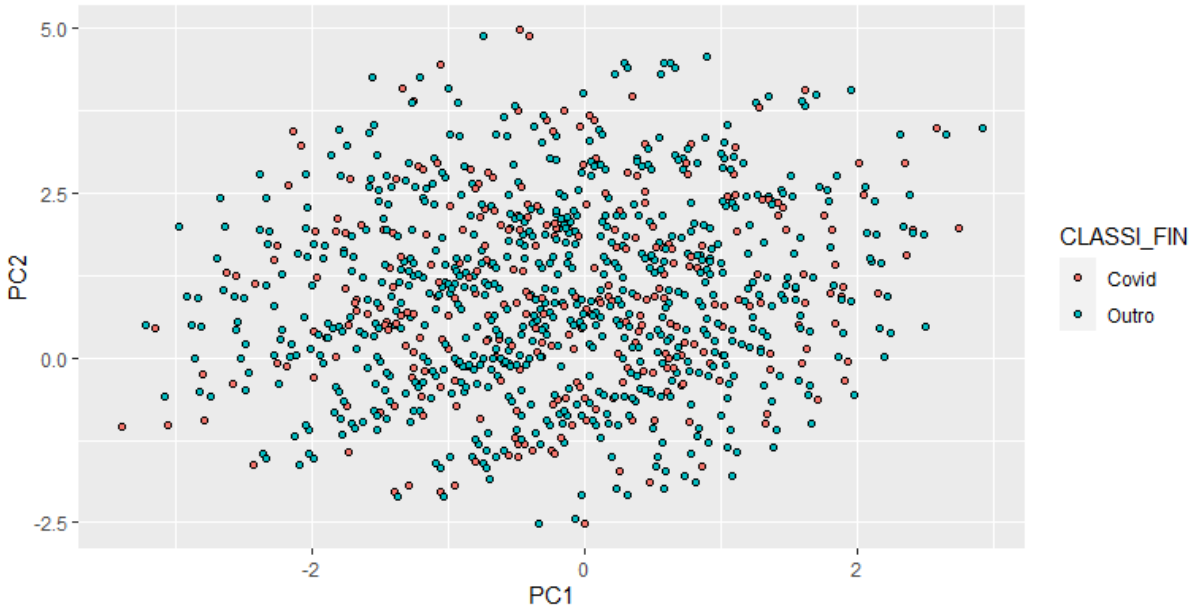


Figura 4.4: Distribuição com duas componentes (29% da variância representada).

Tanto os casos de SRAG Causados por Covid quanto os não causados ficam aglomerados nas mesmas regiões de forma que não é possível distingui-los por suas componentes.

Com isso conseguimos inferir que não é possível distinguir casos de Síndrome Respiratória Aguda Grave causados por Covid pelos causados por outras causas a partir de sintomas e condições pré-existentes.

Capítulo 5

Ajuste de funções aos dados

Uma abordagem fundamental da análise de dados é o ajuste deles por funções matemáticas, pois possuem fácil entendimento, identificam tendências, compatibilidade com a realidade, facilitam a visualização de informações específicas e de relações entre variáveis. Desta forma cada função terá uma combinação de tais atributos favorecendo uns a custo de outros.

Com tais ajustes podemos modelar a propagação do vírus, avaliar e inferir o impacto de intervenções públicas. Deste modo, o primeiro ajuste que contemplamos será o linear, o qual vai ajudar a compreender o ajuste exponencial, que oferece de maneira simples o crescimento inicial do número de infectados quanto outras variáveis da pandemia. Também são utilizadas diversas ferramentas e programas em busca de obter o máximo de informações possível dos dados e explorar os pontos positivos e negativos de cada programa.

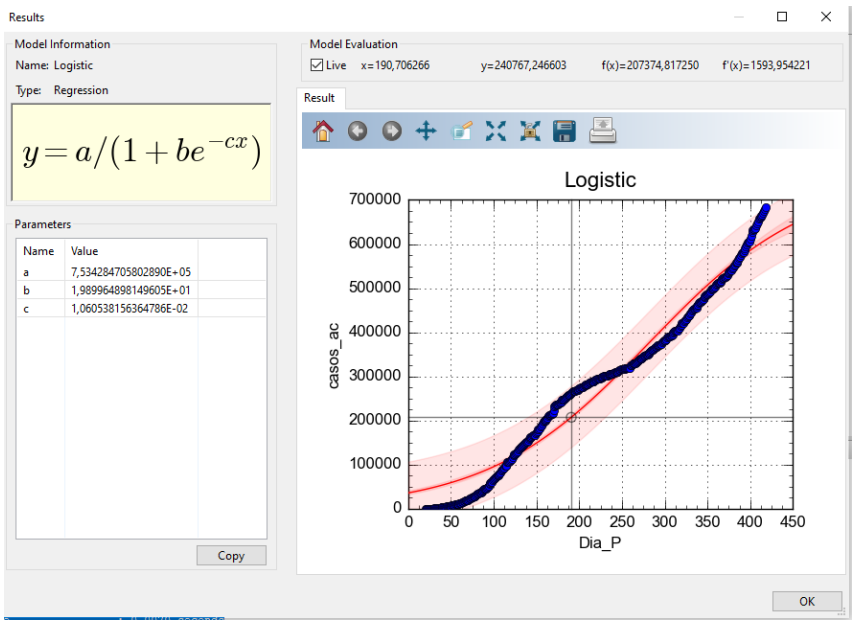


Figura 5.1: Interface do CurveExpert na representação do modelo logistico

5.1 Modelo linear

Antes de analisarmos modelagens mais complexas vamos analisar o ajuste linear, em que mesmo, não tendo uso prático direto nos casos de Covid por sua simplicidade, poderemos usar ele para chegar a ajustes mais complexos e completos. Com este método também podemos calcular o erro de tal ajuste de forma mais compreensível do que no outros casos. Para tal ajuste utilizaremos o método dos mínimos quadrados, tal método fara com que encontremos a função cuja soma dos quadrados dos erros seja mínima, ou seja, sendo os n-ésimos dados observados (x,f(x)) iremos ajustar uma reta g(x) tal que a soma:

$$\sum_{i=1}^{n}(f(x_i) - g(x_i))^2$$

Seja a menor possível. Por ser um ajuste linear temos que a reta g(x) tem forma:

$$g(x) = a_1x + a_0$$

Com isto podemos determinar o erro em função de a_1 e a_0:

$$E(a_1, a_0) = \sum_{i=1}^{n}(f(x_i) - a_1x_i - a_0)^2$$

Para encontrarmos os parâmetros a_1 e a_0 em que $E(a_1, a_0)$ seja mínimo, iremos calcular quando as derivadas parciais de $E(a_1, a_0)$ se anulam, ou seja:

$$\frac{\partial E}{\partial a_1} = 0 \quad \text{e} \quad \frac{\partial E}{\partial a_0} = 0$$

Desta forma os parâmetros a_1 e a_0 são os que satisfazem os seguintes sistemas de equações lineares conhecidas como equações normais:

$$\begin{cases} (\sum_{i=1}^{n} 1)a_0 + \sum_{i=1}^{n} x_ia_1 = \sum_{i=1}^{n} f(x_i) \\ \sum_{i=1}^{n} x_ia_0 + \sum_{i=1}^{n} x_i^2a_1 = \sum_{i=1}^{n} x_if(x_i) \end{cases}$$

Esse sistema pode ser resolvido utilizando qualquer método numérico disponível para solucionar sistemas lineares conhecidos.

Exemplo 5.1. Utilizando os dados da média móvel do número de infectados na cidade de São Paulo podemos utilizar o Excel para criar um gráfico de dispersão e fazer o ajuste linear.

Entretanto, podemos questionar, este é um bom ajuste?

Depende do erro que é a soma dos quadrados dos erros em cada ponto considerado.

	A	B	C	D	E	F	G	H	I	J	K	L
1	nome_mun	codigo_i	datahora	casos	casos	n_sem	obitos	obitos	obitos	pop	semai	casos_mm
564	SÃ£o Paulo	3550308	25/02/2020	1	0		0	0	0	1,2E+07	9	0
1209	SÃ£o Paulo	3550308	26/02/2020	1	0		0	0	0	1,2E+07	9	0
1854	SÃ£o Paulo	3550308	27/02/2020	1	0		0	0	0	1,2E+07	9	0
2499	SÃ£o Paulo	3550308	28/02/2020	2	1		0	0	0	1,2E+07	9	0
3144	SÃ£o Paulo	3550308	29/02/2020	2	0		0	0	0	1,2E+07	9	0
3789	SÃ£o Paulo	3550308	01/03/2020	2	0		0	0	0	1,2E+07	10	0
4434	SÃ£o Paulo	3550308	02/03/2020	2	0		0	0	0	1,2E+07	10	0,142857143
5079	SÃ£o Paulo	3550308	03/03/2020	2	0		0	0	0	1,2E+07	10	0,142857143
5724	SÃ£o Paulo	3550308	04/03/2020	3	1		0	0	0	1,2E+07	10	0,285714286
6369	SÃ£o Paulo	3550308	05/03/2020	6	3		0	0	0	1,2E+07	10	0,714285714
7014	SÃ£o Paulo	3550308	06/03/2020	6	0		0	0	0	1,2E+07	10	0,571428571
7659	SÃ£o Paulo	3550308	07/03/2020	12	6		0	0	0	1,2E+07	10	1,428571429
8304	SÃ£o Paulo	3550308	08/03/2020	15	3		0	0	0	1,2E+07	11	1,857142857
8949	SÃ£o Paulo	3550308	09/03/2020	15	0		0	0	0	1,2E+07	11	1,857142857
9594	SÃ£o Paulo	3550308	10/03/2020	18	3		0	0	0	1,2E+07	11	2,285714286
10239	SÃ£o Paulo	3550308	11/03/2020	29	11		0	0	0	1,2E+07	11	3,714285714
10884	SÃ£o Paulo	3550308	12/03/2020	44	15		0	0	0	1,2E+07	11	5,428571429
11529	SÃ£o Paulo	3550308	13/03/2020	44	0		0	0	0	1,2E+07	11	5,428571429
12174	SÃ£o Paulo	3550308	14/03/2020	62	18		0	0	0	1,2E+07	11	7,142857143
12819	SÃ£o Paulo	3550308	15/03/2020	62	0		0	0	0	1,2E+07	12	6,714285714
13464	SÃ£o Paulo	3550308	16/03/2020	145	83		0	0	0	1,2E+07	12	18,57142857
14109	SÃ£o Paulo	3550308	17/03/2020	156	11		1	1	0,14286	1,2E+07	12	19,71428571
14754	SÃ£o Paulo	3550308	18/03/2020	214	58		3	2	0,42857	1,2E+07	12	26,42857143
15399	SÃ£o Paulo	3550308	19/03/2020	259	45		5	2	0,71429	1,2E+07	12	30,71428571
16044	SÃ£o Paulo	3550308	20/03/2020	306	47		9	4	1,28571	1,2E+07	12	37,42857143
16689	SÃ£o Paulo	3550308	21/03/2020	306	0		9	0	1,28571	1,2E+07	12	34,85714286
17334	SÃ£o Paulo	3550308	22/03/2020	306	0		9	0	1,28571	1,2E+07	13	34,85714286
17979	SÃ£o Paulo	3550308	23/03/2020	306	0		9	0	1,28571	1,2E+07	13	23
18624	SÃ£o Paulo	3550308	24/03/2020	306	0		9	0	1,14286	1,2E+07	13	21,42857143
19269	SÃ£o Paulo	3550308	25/03/2020	722	416		44	35	5,85714	1,2E+07	13	72,57142857
19914	SÃ£o Paulo	3550308	26/03/2020	899	177		53	9	6,85714	1,2E+07	13	91,42857143
20559	SÃ£o Paulo	3550308	27/03/2020	1044	145		62	9	7,57143	1,2E+07	13	105,4285714
21204	SÃ£o Paulo	3550308	28/03/2020	1044	0		62	0	7,57143	1,2E+07	13	105,4285714
21849	SÃ£o Paulo	3550308	29/03/2020	1044	0		62	0	7,57143	1,2E+07	14	105,4285714
22494	SÃ£o Paulo	3550308	30/03/2020	1233	189		103	41	13,4286	1,2E+07	14	132,4285714

Figura 5.2: Distribuição de dados de Covid em São Paulo do Seade - Excel

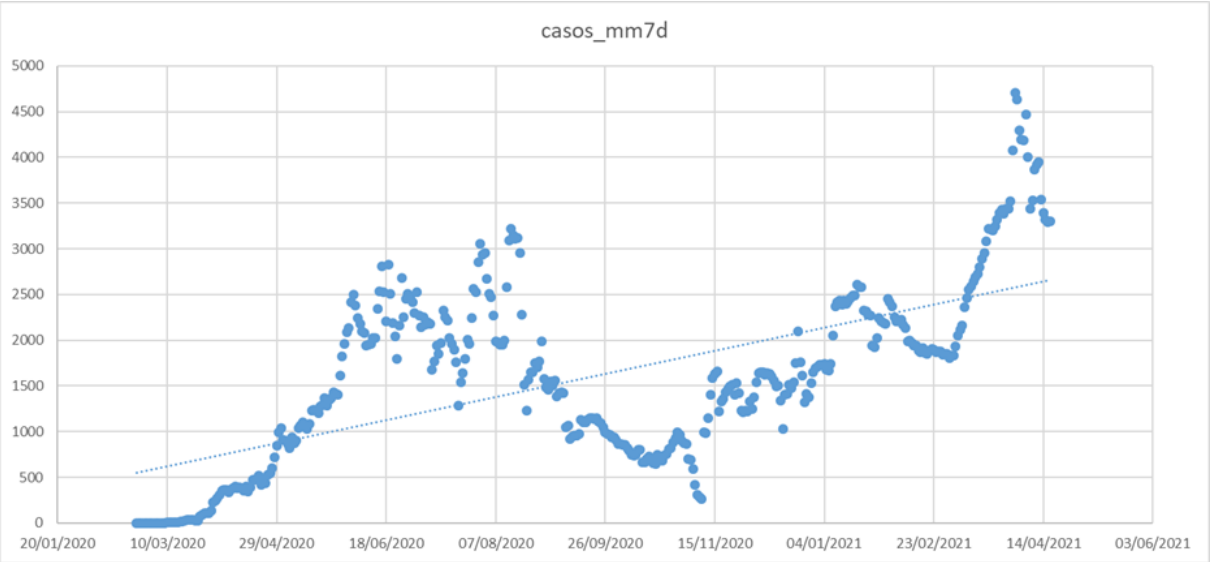
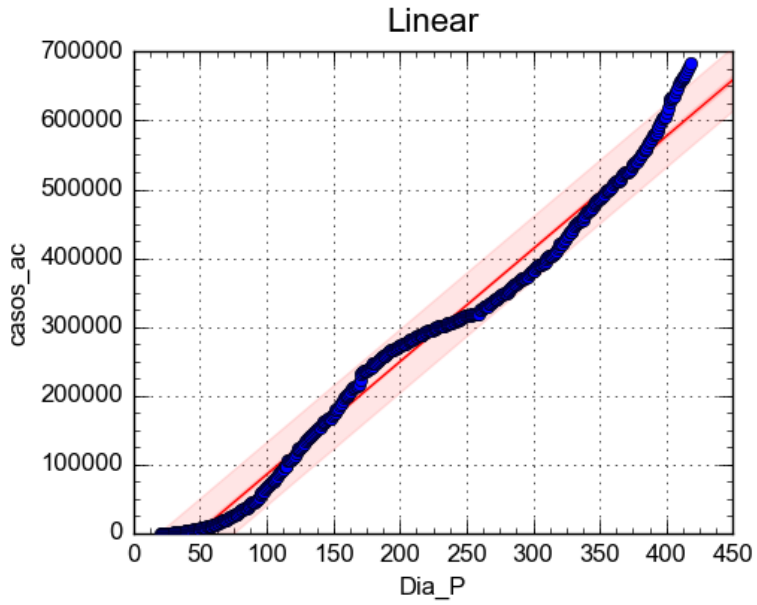


Figura 5.3: Ajuste a Linear a média móvel em São Paulo - Excel



Exemplo 5.2.

Figura 5.4: Ajuste a Linear a casos acumulados em São Paulo - CurveExpert

5.2 Modelo exponencial

Tendo os dados do Seade [5] sobre o número de casos acumulados por dia no estado de São Paulo queremos obter os melhores parâmetros para que o gráfico seja mais próximo da realidade possível, para isto é possível adaptar a função, como se fosse para uma função linear relativa no qual podemos aplicar o mesmo método do modelo anterior. Como queremos modelar utilizando uma função exponencial, então os dados serão aproximados por uma função do tipo:

$$g(x) = ae^{bx}$$

Sendo a,b>0 e x um valor real podemos aplicar o logaritmo natural dos dois lados da equação:

$$z = ln(g(x)) = ln(ae^{bx})$$

e com as propriedades de logaritmo temos:

$$z = ln(a) + bx = b_1\ g_1(x) + b_0\ g_0(x)$$

que é essencialmente um ajuste linear equivalente a um linear dado por:

$$z = b_1 x + b_0$$

ou seja

$$g_1 = x, g_0 = 1 \text{ com } b_1 = b \text{ e } b_0 = ln(a)$$

Formando assim o sistema de equações normais de antes adaptado para:

$$\begin{cases} nb_0 + \sum_{i=1}^{n} x_i b_1 = \sum_{i=1}^{n} ln(f(x_i)) \\ \sum_{i=1}^{n} x_i b_0 + \sum_{i=1}^{n} x_i^2 b_1 = \sum_{i=1}^{n} x_i ln(f(x_i)) \end{cases}$$

Desta forma conseguimos usar métodos numéricos para achar os valores de b_1 e b_0 no qual serão em seguida utilizados par deduzir a e b, já que $g(x) = ae^{bx}$ em que $b = b_1$ e $b_0 = ln(a) \rightarrow a = e^{b_0}$

Exemplo 5.3. Consideramos os dados obtidos dos casos acumulados na cidade de São Paulo, usando do método de relacionar a função exponencial a uma função linear conseguimos obter o ajuste desejado.

	A	B	C	D	E	F	G	H	I	J	K	L
1	nome_mu	codigo_	datahora	casos	casos	n_sem	obitos	obitos	obitos	pop	semar	casos
564	SÃ£o Paulo	3550308	25/02/2020	1	0		0	0	0	1E+07	9	0
1209	SÃ£o Paulo	3550308	26/02/2020	1	0		0	0	0	1E+07	9	0
1854	SÃ£o Paulo	3550308	27/02/2020	1	0		0	0	0	1E+07	9	0
2499	SÃ£o Paulo	3550308	28/02/2020	2	1		0	0	0	1E+07	9	0
3144	SÃ£o Paulo	3550308	29/02/2020	2	0		0	0	0	1E+07	9	0
3789	SÃ£o Paulo	3550308	01/03/2020	2	0		0	0	0	1E+07	10	0
4434	SÃ£o Paulo	3550308	02/03/2020	2	0		0	0	0	1E+07	10	0,1429
5079	SÃ£o Paulo	3550308	03/03/2020	2	0		0	0	0	1E+07	10	0,1429
5724	SÃ£o Paulo	3550308	04/03/2020	3	1		0	0	0	1E+07	10	0,2857
6369	SÃ£o Paulo	3550308	05/03/2020	6	3		0	0	0	1E+07	10	0,7143
7014	SÃ£o Paulo	3550308	06/03/2020	6	0		0	0	0	1E+07	10	0,5714
7659	SÃ£o Paulo	3550308	07/03/2020	12	6		0	0	0	1E+07	10	1,4286
8304	SÃ£o Paulo	3550308	08/03/2020	15	3		0	0	0	1E+07	11	1,8571
8949	SÃ£o Paulo	3550308	09/03/2020	15	0		0	0	0	1E+07	11	1,8571
9594	SÃ£o Paulo	3550308	10/03/2020	18	3		0	0	0	1E+07	11	2,2857
10239	SÃ£o Paulo	3550308	11/03/2020	29	11		0	0	0	1E+07	11	3,7143
10884	SÃ£o Paulo	3550308	12/03/2020	44	15		0	0	0	1E+07	11	5,4286
11529	SÃ£o Paulo	3550308	13/03/2020	44	0		0	0	0	1E+07	11	5,4286
12174	SÃ£o Paulo	3550308	14/03/2020	62	18		0	0	0	1E+07	11	7,1429
12819	SÃ£o Paulo	3550308	15/03/2020	62	0		0	0	0	1E+07	12	6,7143
13464	SÃ£o Paulo	3550308	16/03/2020	145	83		0	0	0	1E+07	12	18,571
14109	SÃ£o Paulo	3550308	17/03/2020	156	11		1	1	0,1429	1E+07	12	19,714
14754	SÃ£o Paulo	3550308	18/03/2020	214	58		3	2	0,4286	1E+07	12	26,429
15399	SÃ£o Paulo	3550308	19/03/2020	259	45		5	2	0,7143	1E+07	12	30,714
16044	SÃ£o Paulo	3550308	20/03/2020	306	47		9	4	1,2857	1E+07	12	37,429
16689	SÃ£o Paulo	3550308	21/03/2020	306	0		9	0	1,2857	1E+07	12	34,857
17334	SÃ£o Paulo	3550308	22/03/2020	306	0		9	0	1,2857	1E+07	13	34,857
17979	SÃ£o Paulo	3550308	23/03/2020	306	0		9	0	1,2857	1E+07	13	23
18624	SÃ£o Paulo	3550308	24/03/2020	306	0		9	0	1,1429	1E+07	13	21,429
19269	SÃ£o Paulo	3550308	25/03/2020	722	416		44	35	5,8571	1E+07	13	72,571

Figura 5.5: Distribuição de dados de Covid em São Paulo do Seade - Excel

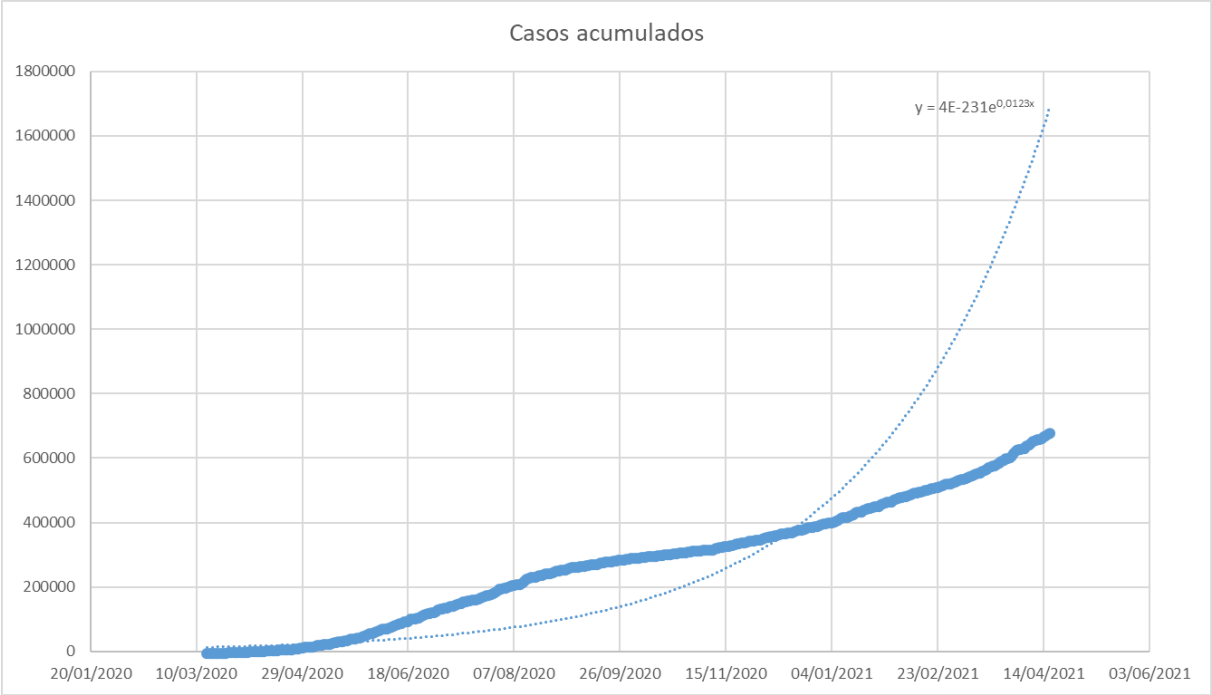
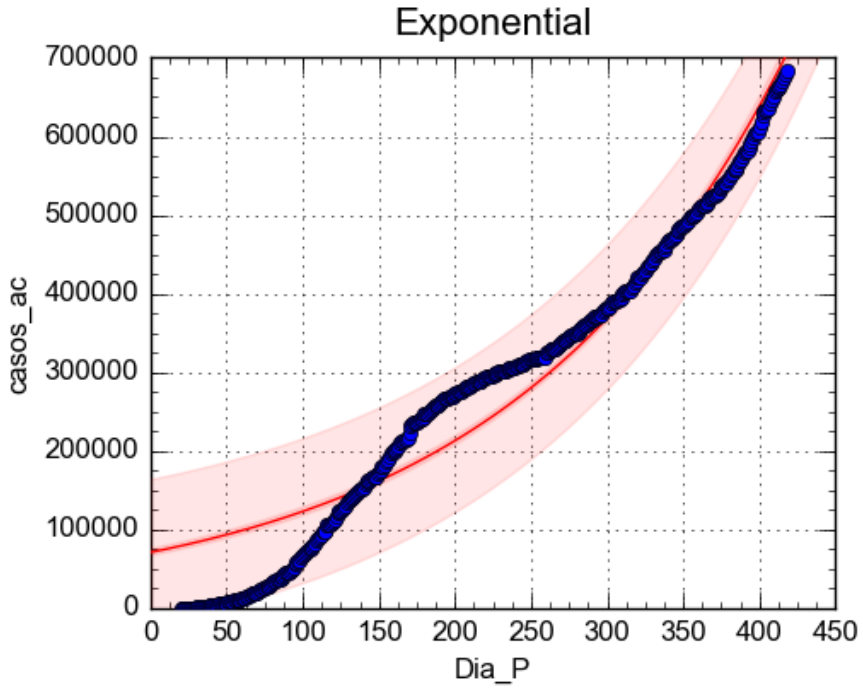


Figura 5.6: Ajuste exponencial feito em casos acumulados em São Paulo - Excel



Exemplo 5.4.

Figura 5.7: Ajuste exponencial a casos acumulados em São Paulo - CurveExpert

5.3 Modelo Logístico

O modelo logístico descreve o crescimento de uma população ou uma quantidade que fica limitada ao longo do tempo. Ele é frequentemente utilizado para modelar situações em que o crescimento é inicialmente rápido, mas diminui à medida que a população se aproxima de um limite superior. O modelo logístico é uma generalização do modelo exponencial, que assume crescimento ilimitado.

Exemplo 5.5.

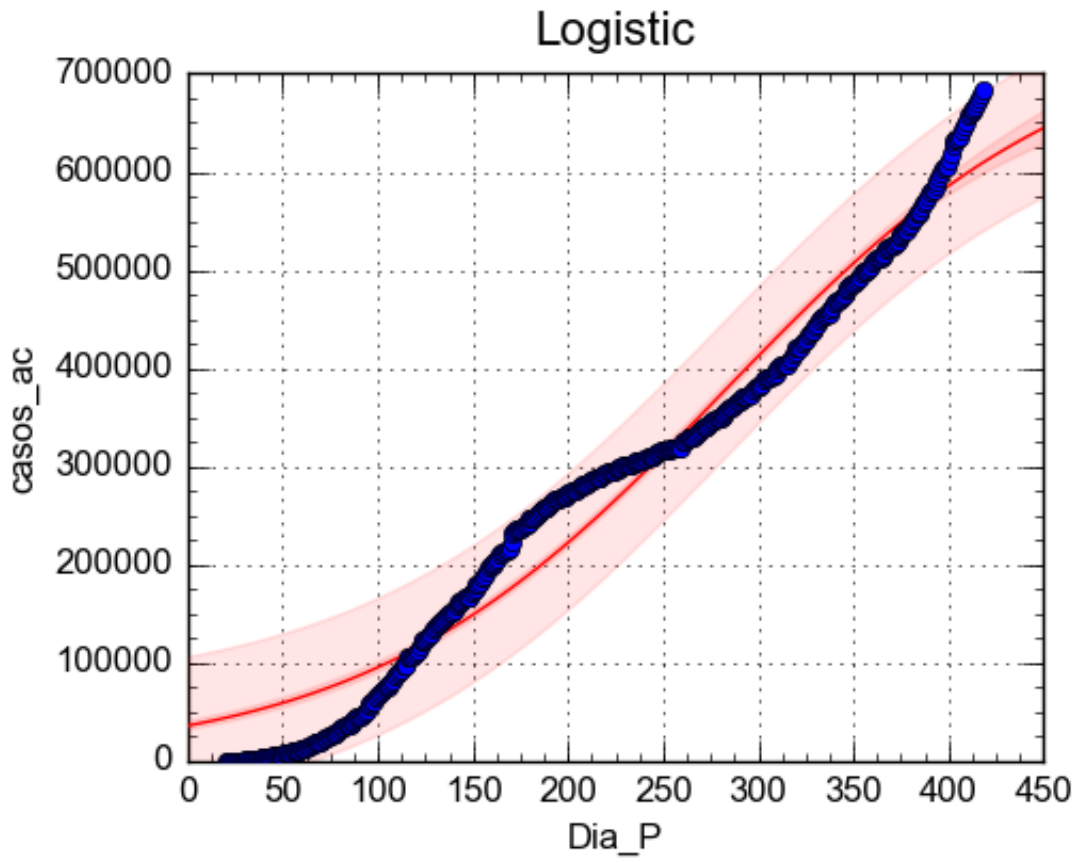


Figura 5.8: Ajuste Logístico a casos acumulados em São Paulo - CurveExpert

Este tipo de ajuste tem um formato sigmoidal calculado por:

$$g(x) = \frac{f_{\max}}{1 + be^{-\lambda x}}$$

em que $f_{\max}$ é o valor máximo da função a que se esta fazendo o ajuste atinge. Suponha agora a transformação:

$$g(x) = \frac{f_{\max}}{1 + be^{-\lambda x}} \approx f(x)$$

podemos notar que existe um ponto de inflexão neste ajuste em:

$$x = \frac{ln(b)}{\lambda}$$

com o valor correspondente temos:

$$f(\frac{ln(b)}{\lambda}) = \frac{f_{\max}}{2}$$

Para estimar um valor razoável para $f_{\max}$ é conveniente escolhermos um valor a cima do ponto de inflexão da função. Podemos agora proceder em isolar parte exponencial da equação, pois fazendo isso é possível trabalharmos com um caso semelhante ao do ajuste exponencial no qual já resolvemos, logo fazendo as seguintes operações:

$$f_{\max} = f(x)(1 + be^{-\lambda x}) \iff \frac{f_{\max}}{f(x)} - 1 = be^{-\lambda x}$$

e ao aplicarmos o logaritmo natural na igualdade temos:

$$ln(\frac{f_{\max}}{f(x)} - 1) = ln(b) - \lambda x = a_1 g_1(x) + a_0 g_0(x)$$

em que

$$a_1 = -\lambda, a_0 = ln(b) \to b = e^{a_0}, g_1(x) = x, g_0(x) = 1$$

Conseguinte, para obtermos os parâmetros a_1 e a_0 devemos resolver o sistema de equações normais dado por

$$\begin{cases} na_0 + \sum_{i=1}^{n} x_i a_1 = \sum_{i=1}^{n} ln(\frac{f_{\max}}{f(x_i)} - 1) \\ \sum_{i=1}^{n} x_i a_0 + \sum_{i=1}^{n} x_i^2 a_1 = \sum_{i=1}^{n} x_i ln(\frac{f_{\max}}{f(x_i)} - 1) \end{cases}$$

Consequentemente, após obtermos a_1 e a_0 podemos encontrar $b = e^{a_0}$ e $\lambda = -a_1$.

Algumas vantagens deste modelo é ser mais realista para descrever o crescimento populacional do que o modelo exponencial. Ele considera que as número de infectados não podem crescer indefinidamente.

No entanto, vemos que o modelo logístico tem suas limitações, especialmente quando se lida com cenários complexos e variáveis externas como nos dados de Covid. Ele pressupõe que as taxas de surgimento e desaparecimento da população em questão são constantes, o que nem sempre é verdade, além disso, o modelo logístico pode não capturar adequadamente flutuações de curto prazo e eventos abruptos.

5.4 Modelo SIR (Susceptíveis-Infectados-Recuperados)

De acordo com a nota técnica divulgada pela UFMG, é possível usar o modelo matemático para demonstrar a situação da quantidade de casos de COVID caso medidas de prevenção não fossem tomadas usando o seguinte sistema:

$$\begin{cases} \frac{dS}{dt} & = \frac{-\beta}{N} IS \\ \frac{dI}{dt} & = \frac{\beta}{N} IS - \gamma I \\ \frac{dR}{dt} & = \gamma I \end{cases}$$

em que S, I e R representam a quantidade de indivíduos suscetíveis, infectados e recuperados e os coeficientes β e γ as taxas de transmissão e recuperação, respectivamente, e N o número de habitantes da região avaliada.

Consideramos a população em 3 compartimentos disjuntos, com o primeiro sendo composto por indivíduos suscetíveis (S), o segundo por indivíduos infectados (I) e o terceiro pelos indivíduos recuperados (R)(também incluindo os falecidos). Temos as seguintes hipóteses para estabelecer o modelo mat:

- Todo indivíduo nasce suscetível;
- O tamanho da população é grande e constante (consideramos a taxa de natalidade igual a taxa de mortalidade e ambas constantes);
- Um indivíduo infectado ganha imunidade;
- A interação entre os indivíduos ocorre de forma homogênea;
- A possibilidade de imigração e emigração são desconsideradas.

Com essas hipóteses, tem-se o objetivo de "simplificar"o modelo em todo seu desenvolvimento, diminuindo as variáveis e facilitando os cálculos.

Ao tentar analisar soluções para o modelo SIR percebemos que as variáveis multiplicadas tornam o sistema não linear, logo tentamos simplificar o sistema transformando as variáveis S e R em produtos de I:

$$I(n+1) - I(n) = \alpha \cdot I(n), \text{ com } \alpha[b, S(n), I(n), R(n)]$$

α: índice de proporcionalidade;
b: taxa de isolamento;
S(n): indivíduos suscetíveis no dia n;

I(n): indivíduos infectados no dia n;
R(n): indivíduos recuperados no dia n.

Portanto, em (1) temos:

$$\begin{aligned} I(n+1) - I(n) &= bS(n)I(n)R(n) \leftrightarrow \\ I(n+1) &= I(n) + bS(n)I(n)R(n) \leftrightarrow \\ I(n+1) &= I(n)[1 + bS(n)R(n)] \end{aligned}$$

Também sabemos que o número total de pessoas (N) é um valor constante e é igual a soma dos termos S(n), I(n) e R(n), que variam no tempo n, ou seja, temos:

$$N = S(n) + I(n) + R(n)$$

Com isso, o próximo passo é discretizar o modelo SIR contínuo em variáveis S(n), I(n) e R(n), como mostra o sistema de equações de diferença a seguir:

$$\begin{cases} S(n+1) - S(n) &= -\frac{\beta}{N} I(n)S(n) \\ I(n+1) - I(n) &= \frac{\beta}{N} I(n)S(n) - \gamma I(n) \\ R(n+1) - R(n) &= \gamma I(n) \end{cases}$$

$$\begin{cases} S(n+1) &= S(n) - \frac{\beta}{N} I(n)S(n) \\ I(n+1) &= I(n) + \frac{\beta}{N} I(n)S(n) - \gamma I(n) \\ R(n+1) &= R(n) + \gamma I(n) \end{cases}$$

Podemos tentar visualizar e simular o modelo SIR, por meio de gráficos no GeoGebra:

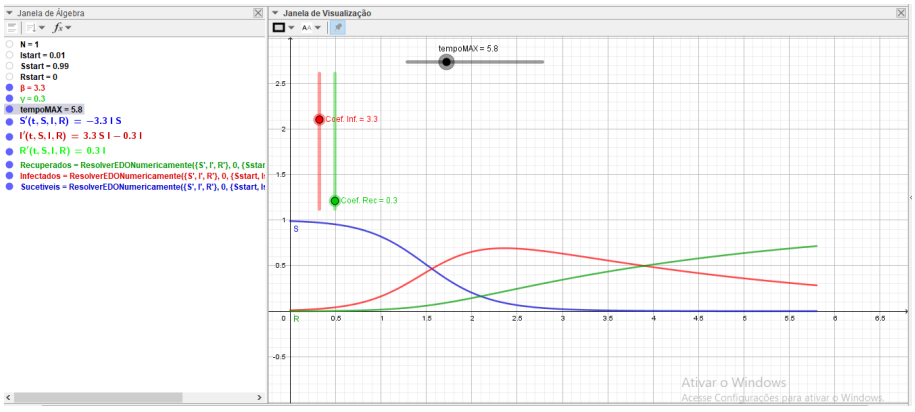


Figura 5.9: Interface do GeoGebra na representação do Modelo SIR

Os controles deslizantes verde e vermelho são utilizados para ajustar os coeficientes β e γ de modo que foi possível perceber que a menor alteração desses coeficientes tem enorme impacto na quantidade de infectados tanto quanto o intervalo de dias que a pandemia durará. Após vários ajustes desses coeficientes e interagindo eles com os dados coletados da cidade São Paulo, foi possível simular um gráfico relativamente realista do início da pandemia.

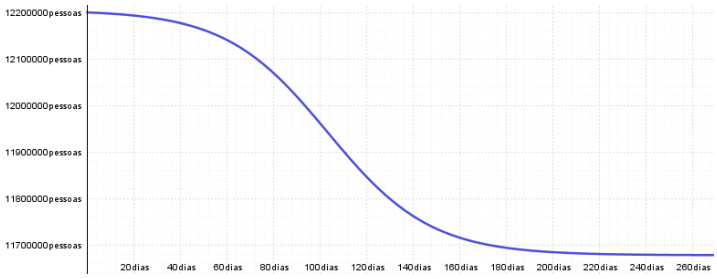


Figura 5.10: Suscetíveis

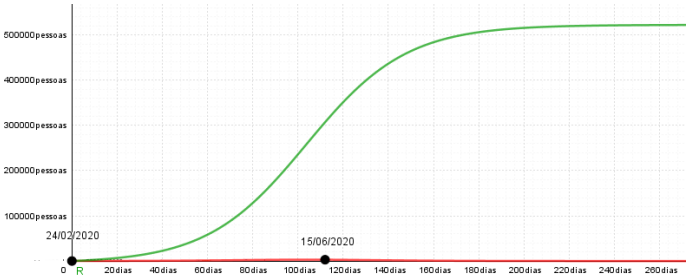


Figura 5.11: Recuperados

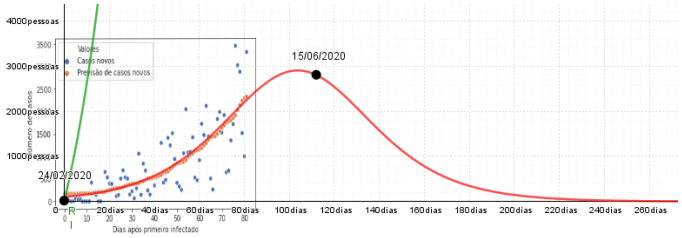


Figura 5.12: Infectados

Modelo SIR Cidade de São Paulo - Geogebra

Utilizando como referência o gráfico feito com a diferença de novos casos minimizada foi possível ajustar os valores dos coeficientes para estimar o comportamento da curva de I, nesse caso temos β=2,105 e γ=2,06 (para ajustar os coeficientes aos dados coletados é preciso aumentar a taxa de recuperação para conseguir uma curva menos alongada horizontalmente e aumentar a taxa de transmissão para aumentar o tamanho geral da curva). Para poder comparar o número de I com os outros foram usadas escalas diferentes para observar suas respectivas curvaturas.

5.5 Modelo SEIR (Susceptíveis-Expostos-Infectados-Recuperados)

O modelo SEIR é uma extensão do modelo SIR que inclui um estágio adicional para os indivíduos expostos (E). Enquanto o modelo SIR considera apenas suscetíveis, infectados e recuperados, o modelo SEIR leva em conta o fato de que após a exposição a uma doença, os indivíduos podem levar algum tempo para se tornarem infecciosos após terem contraído a doença.

O sistema de equações diferenciais do modelo SEIR é dado por:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta \cdot S \cdot \frac{I}{N} \\
\frac{dE}{dt} &= \beta \cdot S \cdot \frac{I}{N} - \alpha \cdot E \\
\frac{dI}{dt} &= \alpha \cdot E - \gamma \cdot I \\
\frac{dR}{dt} &= \gamma \cdot I
\end{aligned}$$

em que:

S é a quantidade de indivíduos suscetíveis

E é a quantidade de indivíduos expostos

I é a quantidade de indivíduos infectados

R é a quantidade de indivíduos recuperados

N é o tamanho total da população

β é a taxa de transmissão da doença

α é a taxa de desenvolvimento dos sintomas

γ é a taxa de recuperação da doença

As equações do modelo SEIR descrevem como os indivíduos se movem entre os grupos ao longo do tempo. Os indivíduos suscetíveis se tornam expostos ao entrar em contato com indivíduos infectados, com uma taxa proporcional a β. Os expostos se tornam infecciosos após um período médio de incubação, determinado por α. Os infectados se recuperam e passam para o grupo de recuperados com uma taxa γ.

O modelo SEIR é utilizado para modelar doenças infecciosas que possuem um período de incubação antes de se tornarem infecciosas. Caso tal período seja relativamente curto, o modelo se aproximara novamente do modelo SIR. Deste modo, ele pode ajudar a prever o impacto da doença em diferentes estágios, a eficácia de medidas de controle e a influência de intervenções, como isolamento e quarentena.

Além das vantagens equivalentes ao do modelo SIR, por levar em consideração o

período de incubação da doença, o modelo é mais realista para doenças que não são imediatamente infecciosas após a exposição, como a gripe e o COVID-19.

Ainda sim, o modelo SEIR também assume uma série de suposições, como a homogeneidade da população e constantes β, α e γ. Modelos ainda mais complexos, podem ser considerados para lidar com cenários mais detalhados e variações maiores dos termos.

5.6 Outros Modelos

Existem diversos outros modelos como, por exemplo, o modelo ARIMA (modelo autorregressivo integrado a média móvel) que analisa séries temporais; Modelos derivados do SIR como o SEIR e SEIRD que trabalham com mais variáveis e parâmetros no intuito de se aproximar da realidade dos dados; modelo Bayesiano que incorporam incertezas e assim fornecendo probabilidades para previsões; Modelo Estocástico que consideram fatores de aleatoriedade representando as flutuações que ocorrem em eventos epidemiológicos e muitos outros.

Capítulo 6

Considerações finais

Comparando os ajustes linear, exponencial e logístico em dados com pequenos intervalos de tempo observamos que não é perceptível uma grande diferença de magnitude quanto aos erros acumulados, porem se aplicarmos os ajustes em um período maior conseguiremos ver uma grande diferença entre os ajustes mais simples comparados com os mais complexos. Com isto, sejam os óbitos acumulados no Brasil a partir do começo de 2020 até o meio de 2022 temos:

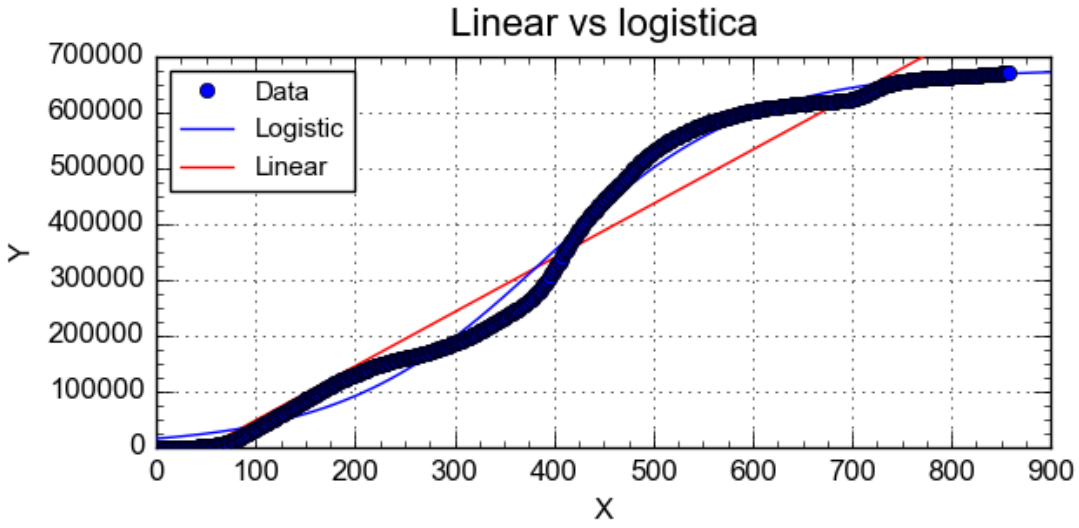


Figura 6.1: Óbitos acumulados no Brasil, dia pandêmico 1 até 900 - CurveExpert
Neste caso o ajuste exponencial teve erro desproporcional e não coube no gráfico.

Vimos como aplicar a análise de componentes principais em dados multivariados e também como podemos aplicar tal técnica em dados de Covid-19 obtidos pelo SEADE. Mesmo que não seja possível identificar o significado de cada componente principal após obtê-las, conseguimos separar os dados para conseguirmos com certa segurança tentar prever a qual categoria um certo elemento se encaixa. Também trabalhamos várias formas diferentes de se manejar e ajustar os dados que podem ser uteis para se tirar conclusões e tomar medidas preventivas contra epidemias.

Referências Bibliográficas

[1] Curso de estatistica , jairo ,6 edição

[2] JOLLIFFE, Ian.T, **Principal Component Analysis**, Segunda edição, Springer, 2002

[3] GRAY, Virginia, **PRINCIPAL COMPONENT ANALYSIS METHODS, APPLICATIONS AND TECHNOLOGY**, New York 2017 by Nova Science Publishers, Inc

[4] https://covid.saude.gov.br/

[5] http://seade.gov.br/coronavirus/

[6] https://ourworldindata.org/coronavirus

[7] MARDIA, K.V.; KENT J.T.;BIBBY J.M., **Multivariate Analysis**,

[8] BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística Básica**, quinta edição.

[9] https://github.com/seade-R/dados-covid-sp

[10] https://rsdjournal.org/index.php/rsd/article/download/23881/20999/285395

[11] KASSAMBARA, Alboukadel, **Practical Guide To Principal Component Methods in R**, primeira edição, STHDA,2017.

[12] SALVADOR, José Antonio; ARENALES, Selma Helena de Vasconselos **Modelagem matemática de problemas ambientais**, coleção UAB - UFSCar,2012.

[13] BURDEM, Richard L.;FAIRES, J.Douglas **Análise numérica**, primeira edição, STHDA,2017.

[14] https://www.curveexpert.net/products/curveexpert-basic/