Análise de dados: uma leitura crítica das informações

Representação de Dados, Ajuste e Correlação

Módulo 2

Gestão da Informação e do Conhecimento

EV.G

Enap

Fundação Escola Nacional de Administração Pública

Diretoria de Desenvolvimento Profissional

Conteudista/s
Ricardo Alexandre Amaral (conteudista, 2022);
Diretoria de Desenvolvimento Profissional.

Enap, 2022
Fundação Escola Nacional de Administração Pública
Diretoria de Desenvolvimento Profissional
SAIS - Área 2-A - 70610-900 — Brasília, DF

Sumário

Apresentação e Boas-vindas

Seja bem-vindo e bem-vinda ao curso Análise de Dados: uma Leitura Crítica das Informações.

Neste curso as noções de estatística serão as lentes fundamentais que nortearão o seu estudo. Mas não se preocupe, as ferramentas estatísticas necessárias para o entendimento do conteúdo serão definidas e retomadas ao longo de todo o curso para ajudar você a alcançar o objetivo geral desta capacitação. Ao final de seus estudos é esperado que você saiba utilizar os fundamentos estatísticos para ler os dados de maneira objetiva e crítica. Dessa forma, o objetivo central do curso se volta à necessidade de entender, na atual era da informação, o conceito de dados e de como tecer uma leitura crítica e objetiva destes.

A estatística é entendida como a ciência dos dados, visto que tem ligação direta com o tratamento desses dados em vários campos do saber. Especialmente neste curso, você poderá analisar o alinhamento da estatística com o desenvolvimento da sociedade, via produção de pesquisa social e por meio de indicadores que permitem mensurar desde características regionais de um país, até suas demandas de desenvolvimento socioeconômico.

O propósito é que este curso sirva para aprimorar um olhar crítico sobre o tema análise de dados e sobre outras áreas de atuação adjacentes ao tema como, por exemplo: gestão, mercado, políticas públicas e a cultura do cidadão pleno. Dessa forma, uma vez que o indivíduo estabeleça vínculo com as mídias informativas espera-se que ele esteja apto à leitura e interpretação de dados.

O conteúdo ao longo do curso está distribuído em três módulos. No primeiro módulo você irá reconhecer qual a importância da análise de dados, irá compreender a conexão entre estatística e tratamento de dados, além de examinar sistemas sob o prisma estatístico.

No segundo módulo você irá reconhecer os principais formatos de representação de dados, analisar correlações não-lineares, além de examinar os ajustes linear e quadrático.

Por fim, no terceiro módulo, irá identificar a contribuição computacional associada ao tratamento de dados, reconhecerá as conexões entre análise de dados e pensamento estratégico, além de reconhecer a importância da análise de dados para a leitura crítica das informações.

Então é hora de começar!

Módulo

2 Representação de Dados, Ajuste e Correlação

O presente módulo do curso tem a proposta de ajudar a ampliar sua criticidade sobre os principais formatos de representação de dados, além disso pretende demonstrar como ocorre a análise de correlações não-lineares e como devem ser examinados os ajustes linear e quadrático.

Logo, o presente módulo abarca a proposta de investigar, por meio de exemplos práticos, didáticos e contextuais, a representação de dados via tabelas, gráficos, ajustes lineares e não-lineares.

Unidade 1: Representação de Dados, Correlação e Ajuste Linear

Objetivo de aprendizagem

Ao final desta unidade você será capaz de reconhecer os principais formatos de representação de dados.

A partir desta premissa você irá aprofundar a sua visão sobre a representação de dados, cujo ajuste pode permitir inferir se há uma regra intrinsecamente relacionada ao conjunto de dados estudado, ou seja, se há uma regularidade que permita prever os pontos, aproximados por uma estrutura matemática conhecida.

1.1 Organizar dados na forma de tabelas e gráficos

Podemos identificar três diferentes “momentos” da elaboração de uma tabela ou gráfico da seguinte maneira:

1) apuração é a forma de aquisição de dados;

2) processamento é a etapa na qual são tratadas tais informações brutas; e

3) apresentação é o modo como serão expostos os dados e os resultados.

Tabelas como formas de organização simples de dados.
Fonte: Freepik. Elaboração: CEPED/UFSC (2022).

Por sua vez, os **gráficos** organizam as informações de modo geométrico, com base na dispersão dos valores, compactação e facilidade de leitura do comportamento global dos dados examinados. Quando o rol de dados é muito extenso, a utilização direta de um gráfico ao invés de uma tabela se torna imprescindível, como por exemplo os gráficos atrelados ao preço de produtos na bolsa de valores ao longo de um ano.

Gráficos como formas de organização de dados de modo geométrico.
Fonte: Freepik. Elaboração: CEPED/UFSC (2022).

1.2 Correlação, linearidade, ajuste linear

É notório que tabelas e gráficos são amplamente difundidos para apresentar dados e ambas apresentam diferentes potencialidades quanto à organização das informações. Órgãos governamentais, educacionais, bancos e jornais, massivamente divulgam dados por meio de tais representações. Tendo essa premissa em vista, conheça alguns conceitos que são fundamentais para a leitura dessas representações:

- **Correlação:** é o grau de conexão entre duas variáveis, qualitativamente entendida como o grau de proximidade dos pontos com alguma função matemática, reta ou parábola etc;

- **Linearidade:** é quando o grau de correlação entre duas quantidades é proporcional e quando dados representados em um gráfico se aproximam intuitivamente de uma reta;

- **Ajuste linear (ou regressão linear):** é o processo de aproximar um conjunto de dados correlacionados por meio de uma reta e a determinação dessa reta que consiga descrever os pontos.

É possível dizer que uma tabela é mais simples, visto que organiza de modo direto os dados, e tem fácil alteração ou conferência, contudo esbarra no problema de ser bastante extensa. Por sua vez, um gráfico é altamente compacto, mostra o comportamento global dos dados e tendências, assim como o crescimento e decrescimento máximo e mínimo etc.

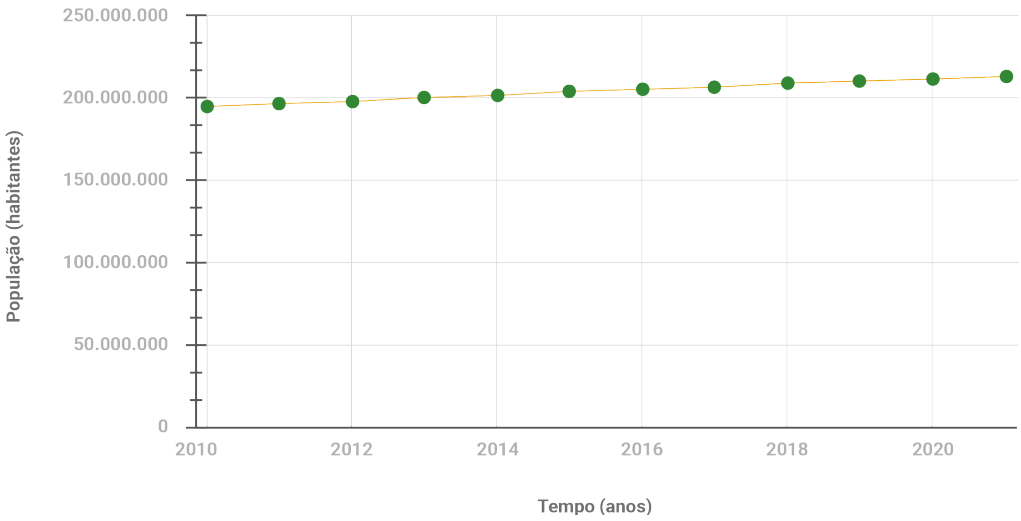
A tabela a seguir, que usa como base as informações do IBGE, sumariza números da população absoluta do Brasil no período de 2010-2021. É importante notar que tal representação é direta, simples e não evidencia comportamento geral da informação, como máximos, mínimos, tendências etc.

Período (anos)	População (milhões de habitantes)
2010	194,9
2011	196,6
2012	198,3
2013	200
2014	201,7
2015	203,5
2016	202,2
2017	206,9
2018	208,5
2019	210,2
2020	211,8
2021	213,4

Dados da População Brasileira Absoluta no Período 2010-2021.

Fonte: IBGE (2021). Elaboração: CEPED/UFSC (2022).

O gráfico abaixo está diretamente relacionado aos dados da tabela anteriormente apresentada. Na verdade, são os mesmos dados rearranjados sob outra apresentação gráfica. A leitura de um gráfico oferece inúmeras vantagens, como a grande compactação das informações, facilidade de entendimento global do conjunto de dados, tal como pode ser visto na curva que atesta um crescimento da população ao longo do período. Há vários modelos de gráficos e no exemplo em particular está representado um modelo que é denominado gráfico de linhas de dispersão. Esse tipo de representação basicamente organiza em um plano cartesiano os dados antes estruturados na forma de tabela.



Dados da População Brasileira Absoluta no Período 2010-2021 em representação gráfica.
Fonte: IBGE (2021). Elaboração: CEPED/UFSC (2022).

Sobre o gráfico e a tabela anteriores é importante citar que:

- são modos distintos de apresentar o mesmo conjunto de dados;
- os pontos azuis são as interseções dos dados dispostos na tabela anterior, como por exemplo para o primeiro ponto, 2010 e 194,9 milhões, ou (2010 ;194.900.000), temos o cruzamento horizontal e vertical de tais dados, e assim sucessivamente;
- as linhas escuras horizontal e vertical são chamadas de **eixos** e estão respectivamente relacionadas aos rótulos das informações e ao posicionamento dos dados, neste caso tempo e população;
- o fundo quadriculado é apenas para orientação das informações espacialmente distribuídas no gráfico, visto que comumente ajuda na leitura dos pontos, todavia podendo ser omitido sem problemas;

- a linha roxa liga os pontos no gráfico e é muito importante para o estudo, visto que representa o tipo de ajuste dos dados;

- um exame visual intuitivo da distribuição dos dados no gráfico sugere que tais pontos exibem uma tendência de reta (correlação), em outras palavras, que uma dada reta poderia passar por todos pontos, ou ainda que é possível ajustar de modo aproximado (regressão linear) o rol de pontos por uma reta.

Encontrar a reta de ajuste é o que constitui o processo de **regressão linear**, que pode ser manual ou feito com uso de recursos computacionais. A correlação, por sua vez, é um parâmetro que mede a fidelidade do ajuste feito, ou seja, investiga se a regra matemática escolhida é suficientemente boa.

A intenção de se estudar a regressão linear é que ela retrata uma regra simples entre duas quantidades e permite inferências futuras.

Desta maneira, fazer uma regressão linear das informações acerca da população brasileira em um determinado período significa procurar estimar se uma quantidade de cidadãos se altera no tempo observado e se há aproximação de tais informações por uma reta.

O resultado disso é o uso de regras matemáticas que definem uma reta para realizar projeções sobre a evolução destes dados ou sobre tendências de tal sistema em um futuro próximo.

DESTAQUE

Conceitualmente é possível entender que o ajuste, regressão linear, ou *linear fit* é um mecanismo de determinar uma função que descreve um conjunto de dados.

A correlação indica quão confiável é a aproximação usada.

É importante ressaltar que modelos de tratamentos de dados, ou seja, ajustes, como o de **regressão linear**, são sempre aproximados, devido à natureza da medida, que tem um erro intrínseco dada a não existência de processos nem mecanismos perfeitos de coleta de dados. Deste modo, uma regressão é uma aproximação, um ajuste de um rol de dados para uma forma matemática pré-existente, tal como uma reta, uma parábola etc, que nunca será perfeito.

O processo de regressão linear é feito em observação à equação da reta, ajustando-se, a partir da realidade dos dados, os parâmetros de uma reta. O resultado é uma equação matemática que descreve aproximadamente a dependência funcional entre as variáveis. Deste modo, com o uso da equação da reta é possível construir a equação e assim desvendar um modelo aproximativo que descreve os pontos, permitindo realizar uma projeção dos resultados futuros.

Para isso é utilizada a equação reduzida da reta, comumente escrita como:

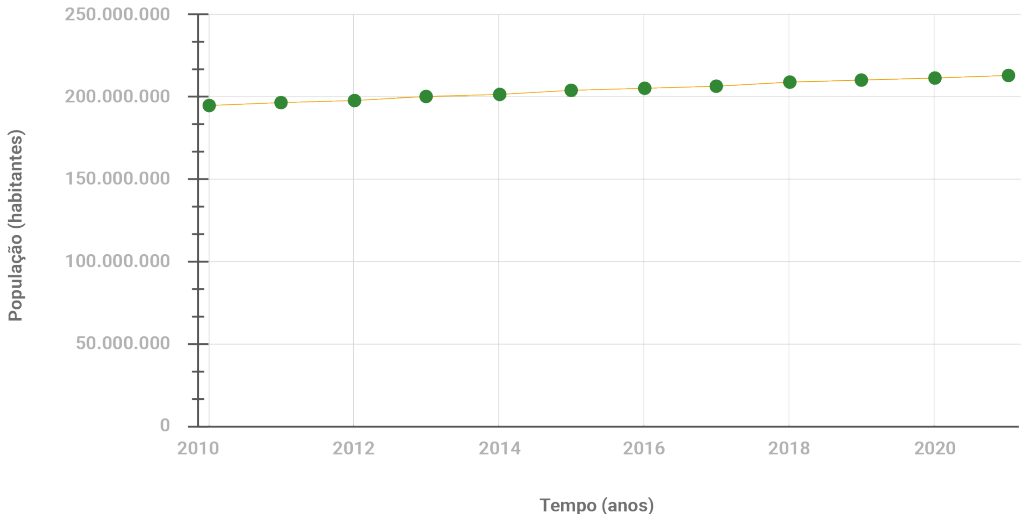
$$y = Ax + B$$

A fórmula mostra a dependência entre as variáveis, **x e y**, que são respectivamente a **variável independente** e a **variável dependente**.
A é o **coeficiente angular** responsável pelo crescimento da equação **e B** o **coeficiente linear** "onde começa a equação".

Uma reta descrita pela equação y = 2x + 4 significa que para cada x dado há uma correspondência em y, o qual se inicia em 4 (quando x=0) e aumenta ao passo 2.

Assim, no estudo dos dados da evolução da população brasileira com relação ao tempo, que tem um recorte representado no gráfico e na tabela anteriores, é possível, utilizando a grande correlação dos pontos com uma reta, determinar os coeficientes angulares e lineares, e assim determinar uma reta pela qual seja possível estimar a evolução temporal do número de brasileiros.

Reveja aqui o gráfico para que possa ser mais simples compreender, a seguir, alguns conceitos.



Dados da População Brasileira Absoluta no Período 2010-2021 em representação gráfica.
Fonte: IBGE (2021). Elaboração: CEPED/UFSC (2022).

Vale lembrar que este gráfico se baseia em dados de números da população absoluta do Brasil no período de 2010-2021 que estão organizados em tabela demonstrada anteriormente, iniciando em 2010 com a população de 194,5 milhões de habitantes e finalizando em 2021 com a população de 213,4 de habitantes.
Neste sentido, para determinação dos parâmetros que regem a reta representada no gráfico se procede da seguinte forma:

- o primeiro e último ponto representam o coeficiente linear B (é o ponto que corta o eixo vertical em aproximadamente 194,5 milhões), enquanto o coeficiente angular é a variação vertical com relação à horizontal, matematicamente definida pela razão de tais alterações;

- com relação à evolução da população brasileira, que se comporta como reta, é possível estudar dois pontos quaisquer dentro da amostra, e para o presente caso foram analisados o primeiro e último ponto destes dados

- B é o ponto que corta o eixo vertical em aproximadamente 194,5 milhões, logo o coeficiente linear é representado como B = 194,5 milhões;

- o coeficiente angular é a variação vertical, diferença entre o valor do coeficiente angular e a população em 2021, que se traduz no cálculo 213,4 milhões - 194,5 milhões = 18,9 milhões, com relação a horizontal 2021 - 2010 = 11, matematicamente definida pela razão de tais alterações, e que se relaciona como uma velocidade de crescimento ou decrescimento, no caso determina com qual velocidade a população cresce, ou decresce, no período examinado;

- logo o coeficiente angular é a variação vertical, razão de tais alterações, e que se relaciona no seguinte cálculo: (213,4 milhões - 194,5 milhões)/ (2021 - 2010) = 18,9 milhões/11 = 1,71 milhão por ano.

- o resultado é que se torna possível escrever uma equação que ajusta o conjunto de dados com relação ao crescimento da população dada por uma reta, genericamente y = Ax + B, cujos parâmetros são A=1,71 milhão/por ano e B= 194,5 milhões.

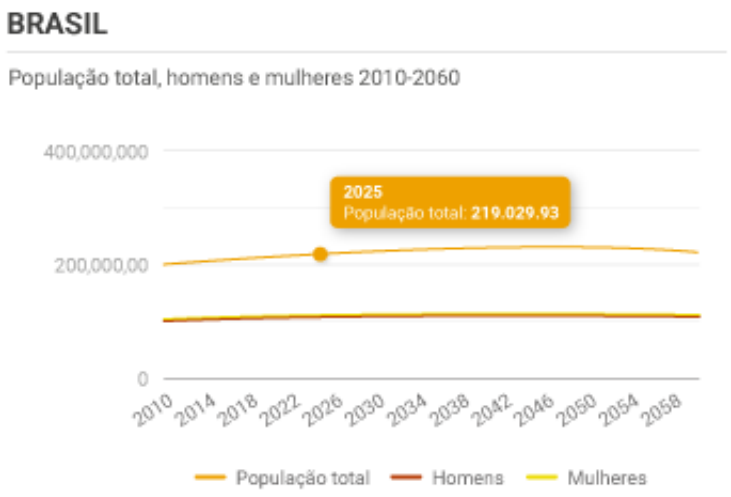
Utilizando todos esses resultados é alcançada a seguinte relação na equação:

$$y = 1{,}71x + 194{,}5.$$

A equação acima traduz que o ponto de partida ocorreu em y = 194,5 milhões para tempo zero, isto é, 2010, e a cada ano temos um acréscimo de 1,71 milhão.

A aproximação linear é muito interessante e importante principalmente para estimativas rápidas e não muito extensas na variável independente "x", ou seja, projeções de curto prazo. Perceba que essa mesma equação promove um resultado muito próximo dos modelos mais sofisticados para o ano de 2025, ou seja, quando estuda uma pequena variação do rol de dados analisados e ajustados.

De modo particular podemos entender o ano inicial como 2010 ou x = 0 e 2025 como o ano x = 15 com base no intervalo de tempo entre tais marcos, e assim utilizar a equação de ajuste para calcular uma estimativa para a população em 2025, como y = 1,71(15) + 194,5 = 220 milhões, o que está de acordo com as previsões do IBGE para a população brasileira, como mostra o gráfico abaixo.



População total brasileira, homens e mulheres 2010-2060.
Fonte: IBGE (2021). Elaboração: CEPED/UFSC (2022).

Consequentemente a regressão neste caso é amplamente satisfatória, visto que é um recurso simples, direto e consegue estimar corretamente a evolução de sistema, como neste caso o crescimento da população brasileira em um determinado período. A finalidade da regressão linear é justamente essa, conseguir uma forma de fazer previsões. A figura acima mostra uma curva de projeção da população brasileira mais realística, baseada em modelos mais sofisticados, associada ao período de 2010-2060, além de destacar a previsão para 2025.

Você chegou ao fim desta unidade. Parabéns! Caso surjam dúvidas sobre o tema, reveja o conteúdo para fixar seus estudos.

Referências

ALVES, Isabel Fraga. Data Science, Big Data e um novo olhar sobre a Estatística. **Boletim SPE**: O Tema Central da Estatística - um novo olhar, Lisboa, v. 12, n. 2, p. 29-31, 2017. Semestral.

CARVALHO, Marilia Sá; SOUZA-SANTOS, Reinaldo. Análise de dados espaciais em saúde pública: métodos, problemas, perspectivas. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 21, p. 361-378, 2005.

HURWITZ, Judith et al. **Big Data para leigos**. Rio de Janeiro, Alta Books Editora, 2016.

INSTITUTO BRASILEIRO DE GEOGRAFIA (IBGE). **O que é o PIB**. Rio de Janeiro, IBGE, 2021. Disponível em: https://www.ibge.gov.br/explica/pib.php. Acesso em: 12 nov. 2021.

LEITE, Angela; FUITA, Hiroco. **Aplicações da matemática: administração, economia e ciências contábeis**. Boston, Cengage Learning, 2008.

PINTO, José Carlos; SCHWAAB, Marcio. **Análise de Dados Experimentais: I. Fundamentos de Estatística e Estimação de Parâmetros**. Rio de Janeiro, Editora E-papers, 2007.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business**: What you need to know about data mining and data-analytic thinking. Sebastopol (USA) O'Reilly Media, Inc., 2013.

SILVESTRE, António. **Análise de dados e estatística descritiva**. Forte da Casa, Escolar Editora, 2007.

VITALI, Marieli Mezari. Estatística sem matemática para psicologia. **Revista Brasileira de Psicodrama**, São Paulo, v. 27, n. 1, p. 139-144, 2019.

VUOLO, José Henrique. **Fundamentos da teoria de erros**. São Paulo, Editora Blucher, 1996.

Unidade 2: Debater o Comportamento Não-linear via Exemplos

Objetivo de aprendizagem

Ao fim desta unidade você será capaz de analisar correlações não-lineares e ampliar os conceitos sobre ajuste de dados.

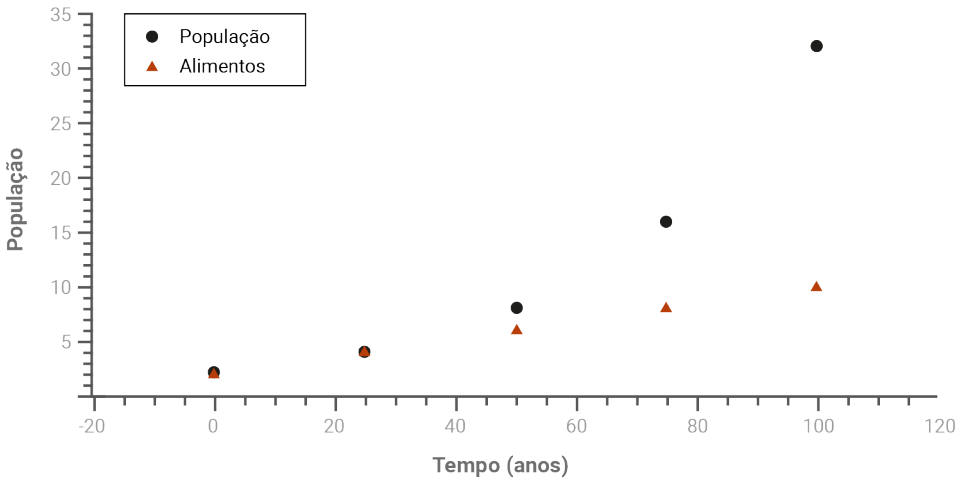
A regressão linear é um processo que busca uma reta $y = Ax + B$ que mais se aproxima dos pontos representados por meio de um gráfico, sendo que os dados submetidos satisfatoriamente a tal aproximação por reta são entendidos como lineares. Na mesma perspectiva, a regressão não-linear, é uma aproximação dos dados "x e y" utilizando uma estrutura matemática conhecida que não seja uma reta (linear). A parábola (equação de grau-2, $y = Ax^2 + Bx + C$) é um bom exemplo de estrutura desse tipo que se presta também para descrição da fenomenologia dos dados.

A **linearidade** de um fenômeno é característica muito apreciada, visto que tem uma matemática simples diretamente atrelada à equação da reta. Contudo, ao se ajustar dados provenientes de diversas áreas, comumente evidencia-se outros tipos de comportamento discrepantes de uma reta, o que promove a busca de funções matemáticas com melhor ajuste com relação a esses dados.

Uma reta varia sempre da mesma forma, isto é, cresce ou decresce de forma regular, enquanto isso, a não-linearidade quebra tal característica e se caracteriza por variações que não se mantém.

A teoria demográfica de Malthus do final do século XVIII, também conhecida como Lei de Malthus, afirmava que as populações humanas iriam se duplicar a cada 25 anos caso não ocorressem guerras e pestes. Cresceria, portanto, em progressão geométrica 2, 4, 8, 16, 32, não-linear. Em contrapartida a produção de alimentos aumentaria apenas em progressão aritmética, 2, 4, 6, 8, 10, linear.

Embora, felizmente, os resultados malthusianos não tenham se comprovado, já que não consideravam os avanços agrícolas, o exemplo é muito interessante para o propósito de comparar linearidade e não-linearidade, assim como também é possível observar no gráfico a seguir.



Visualizando a Lei de Malthus, século XVII.
Fonte: Elaboração do autor. Elaboração: CEPED/UFSC (2022).

Mentalmente é possível traçar uma reta que une os pontos triangulares da figura acima e que descrevem o aumento da produção de alimentos, ou seja, é possível ajustar a curva de crescimento de alimentos como uma reta, ao passo que o mesmo seria impraticável para a curva populacional representada pelos pontos em preto. No exemplo é perceptível que a reta cresce sempre da mesma forma, o que não ocorre para a população, não-linear segundo o modelo malthusiano.

A não-linearidade dos dados é frequente nas mais vastas áreas de conhecimento, existem inúmeras funções matemáticas que podem servir para tais ajustes como logaritmos, exponenciais, polinômios etc. O foco será abordar o ajuste quadrático (polinomial de grau-2), o qual se trata da equação da parábola, que representa ou utiliza a estrutura de uma função quadrática para expressar um conjunto de dados.

2.1 Tipos de comportamentos não-lineares e estudo da parábola

Ao tecer um gráfico é muito comum que uma curva estabeleça um comportamento mais próximo de funções mais gerais do que uma reta, ou que curvas como a parábola melhor se moldem com os dados analisados.

DESTAQUE

O problema dos ajustes não-lineares é atrelado a abrupto aumento de dificuldade, pois enquanto muitos ajustes lineares podem ser feitos manualmente, os ajustes não-lineares, mesmo os mais simples, já se tornam terrivelmente enfadonhos de calcular manualmente e comumente são executados via uso de recursos computacionais.

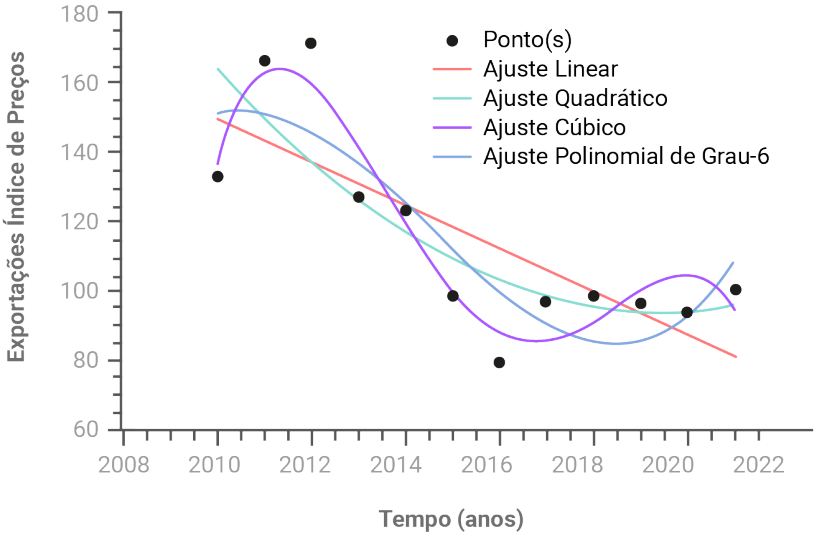
Uma forma simples para entender a curva malthusiana é normalizar o tempo, isto é, pensar que a cada 25 anos há uma unidade de tempo, como na relação t' = 25, ou seja, os alimentos crescem sob a forma 2, 4, 6, 8, 10 o que sugere uma função y = 2 + 2t' (função linear que relaciona t' e y, que se inicia em 2 e cresce de 2 em 2).

Ao contrário, a curva da população tem um comportamento mais complexo, o qual não se ajusta como reta, dado que não mantém sua variação, o qual deve ser ajustado a funções de ordens mais altas como a quadrática, y = At'2 + Bt' + C, sendo A, B e C parâmetros a serem determinados.

2.2 Construção numérica de ajustes *(fit)* e sua interpretação

O gráfico a seguir mostra um conjunto de dados colhidos do site IPEADATA, relacionados ao índice de preços das exportações, e foram submetidos a diferentes ajustes computacionais lineares e não-lineares.

A execução de um ajuste tem sempre a mesma lógica de aproximar um conjunto de pontos por uma função matemática similar à disposição de pontos previamente conhecida. Basicamente aqui serão vistas as regressões linear e quadrática, dados que se comportam e não se comportam como reta, respectivamente.



Índice de preços das exportações: total geral.
Fonte: IPEADATA (2021). Elaboração: CEPED/UFSC (2020).

Os ajustes estão representados de acordo com a legenda, sendo regressão linear em vermelho (reta), quadrático em azul claro, cúbico em azul escuro e de sexto grau em roxo.

Note que o ajuste linear tem uma correlação bastante insatisfatória visto que não se aproxima da maioria dos pontos, os ajustes não-lineares, ao contrário, promovem uma melhor correlação, dado que conseguem melhor se moldar ao conjunto de pontos, todavia a um custo maior de complexidade matemática.

É válido dizer que o ajuste linear tem seu crescimento comparado a uma regra de três, tal como a estimativa feita aqui para o crescimento de 1,71 milhão/ano para a população brasileira em um curto intervalo.

Você chegou ao fim desta unidade. Caso surjam dúvidas, faça uma releitura dos tópicos de interesse. Bons estudos!

Referências

ALVES, Isabel Fraga. Data Science, Big Data e um novo olhar sobre a Estatística. **Boletim SPE**: O Tema Central da Estatística - um novo olhar, Lisboa, v. 12, n. 2, p. 29-31, 2017. Semestral.

CARVALHO, Marilia Sá; SOUZA-SANTOS, Reinaldo. Análise de dados espaciais em saúde pública: métodos, problemas, perspectivas. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 21, p. 361-378, 2005.

COMARELA, Giovanni et al. Introdução à Ciência de Dados: Uma Visão Pragmática utilizando Python, Aplicações e Oportunidades em Redes de Computadores. SCHAEFFER FILHO, Alberto Egon; CORDEIRO, Weverton Luis da Costa; CAMPISTA, Miguel Elias Mitre (ed.). **Minicursos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**, Porto Alegre, Sociedade Brasileira de Computação, p. 246-295, 2019.

GIBBS, Graham. **Análise de dados qualitativos**. Porto Alegre, Artmed; 2009.

HURWITZ, Judith et al. **Big Data para leigos**. Rio de Janeiro, Alta Books Editora, 2016.

INSTITUTO BRASILEIRO DE GEOGRAFIA (IBGE). **O que é o PIB**. Rio de Janeiro, IBGE, 2021. Disponível em: https://www.ibge.gov.br/explica/pib.php. Acesso em: 12 nov. 2021.

LEITE, Angela; FUITA, Hiroco. **Aplicações da matemática: administração, economia e ciências contábeis**. Boston, Cengage Learning, 2008.

PINTO, José Carlos; SCHWAAB, Marcio. **Análise de Dados Experimentais: I. Fundamentos de Estatística e Estimação de Parâmetros**. Rio de Janeiro, Editora E-papers, 2007.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business**: What you need to know about data mining and data-analytic thinking. Sebastopol (USA) O'Reilly Media, Inc., 2013.

SILVESTRE, António. **Análise de dados e estatística descritiva**. Forte da Casa, Escolar Editora, 2007.

VITALI, Marieli Mezari. Estatística sem matemática para psicologia. **Revista Brasileira de Psicodrama**, São Paulo, v. 27, n. 1, p. 139-144, 2019.

VUOLO, José Henrique. **Fundamentos da teoria de erros**. São Paulo, Editora Blucher, 1996.

Unidade 3: Praticar os Conceitos de Ajuste de Curvas

Objetivo de aprendizagem

Ao final desta unidade você será capaz de examinar os mecanismos de ajustes linear ou quadrático executados computacionalmente.

Veja a seguir os processos embarcados nos ajustes computacionais, desde o raciocínio necessário para aproximar dados às funções, até a compreensão de modelos com base nos ajustes.

3.1 Prática de *fit*: principais ajustes, aproximações e tendência

Neste momento do conteúdo busca-se examinar os mecanismos alinhados aos ajustes linear e quadrático, que são executados de forma mais prática usando recursos computacionais, isto é, programas que executam de modo ágil praticamente todas as etapas de uma regressão.

Os métodos computacionais são vastamente difundidos no ajuste de curvas de dados, pois realizam um árduo trabalho automatizando diversas operações e entregando os ajustes prontos.

Para que esse processo ocorra os passos são os seguintes:

- desenhar uma curva;
- observar com qual tipo de função ela se parece (aqui restritos a reta ou parábola);
- pedir que o computador faça o *fit* de tal curva.

Em conjunto com este processo serão utilizados no conteúdo ensaios de programas de computador com a finalidade de desenvolver um olhar computacional com relação às ferramentas básicas de análise de dados.

Há inúmeros programas para tal finalidade, cuja implementação pode ser definida com relação ao volume de trabalho computacional. Um exemplo são os sistemas

altamente automatizados de dados, outros são aqueles com apuração humana e que promovem tratamento comumente feito por planilhas eletrônicas e/ou programas estatísticos especializados.

Serão usadas as linguagens de programação Python e R. A linguagem Python é flexível e enxuta; enquanto R é dedicada à análise estatística, portanto muito do código já está pronto para uso estatístico.

Dentre a variedade de *softwares* possíveis para ajuste computacional de gráficos serão abordados aqui dois sistemas com foco na produtividade e didática: uma planilha eletrônica (Planilhas Google) e o SciDAVis.

DESTAQUE

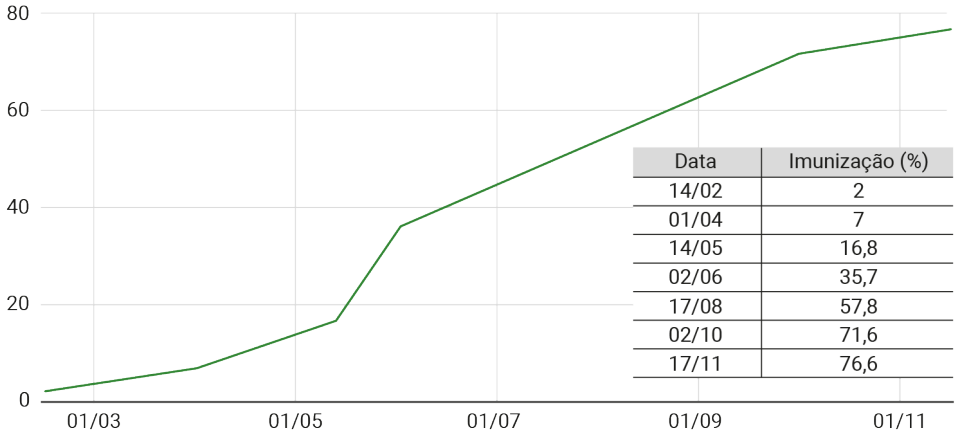
O nome SciDAVis é na verdade um acrônimo para Scientific Data Analysis and Visualization, sendo um programa de computador de código aberto, livre, amigável, multiplataforma (Linux, Windows, Mac) e que possui ampla documentação disponível na internet.

3.2 Leitura de gráficos

A experiência que se propõe aqui é de ler um gráfico de modo sistêmico, isto é, verificar os pontos, suas tendências e possíveis ajustes, e com isso tirar conclusões técnicas acerca deste.

A figura a seguir mostra a confecção de um gráfico em uma planilha eletrônica convencional (Planilhas Google) e cujos dados representam a curva do percentual de imunização parcial com a primeira dose da vacina contra o coronavírus no Brasil, por meio dos dados disponibilizados no site Our World in Data (2021).

Curiosamente tais planilhas não mostram os pontos discretos, ao contrário, apresentam uma linha que pode confundir os dados como contínuos, e por consequência gerar certa ambiguidade.

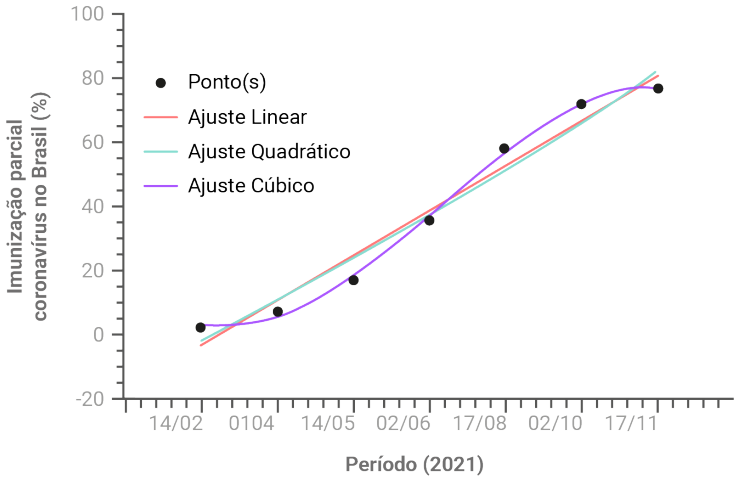


Data	Imunização (%)
14/02	2
01/04	7
14/05	16,8
02/06	35,7
17/08	57,8
02/10	71,6
17/11	76,6

Dados e esboço da curva do percentual de imunização parcial com a primeira dose da vacina contra o coronavírus no Brasil ultilizando a Planilha Google.

Fonte: Our Wold in Data (2021). Elaboração: CEPED/UFSC (2022).

A seguir uma figura do refino da curva de imunização parcial com a primeira dose da vacina contra o coronavírus no Brasil, utilizando o SciDAVis, que alia qualidade tipográfica em relação a uma planilha genérica, como também facilidade no ajuste de dados (linhas coloridas). A confecção do gráfico foi feita introduzindo os dados, selecionando-os e com uso dos menus: *Plot* (do verbo plotar, esboçar) e em seguida *Scatter* (dispersão, discreto, separado), resultando no gráfico a seguir.



Dados e esboço da curva do percentual de imunização parcial com a primeira dose da vacina contra o coronavírus no Brasil ultilizando o SciDAVis.

Fonte: Our Wold in Data (2021). Elaboração: CEPED/UFSC (2022).

Os ajustes mostrados na curva anterior sobre imunização parcial no Brasil são representados por linhas contínuas na figura. Tendo em vista a figura, legenda e os resultados é interessante discutir os seguintes pontos em relação ao contexto e à notação computacional/matemática:

- **Ponto(s):** são os pontos relativos à tabela que deu origem ao gráfico;
- **Ajuste Linear:** linha vermelha, relativa à regressão linear (reta, grau-1), aproximada por uma função do tipo y = A*x + B (o asterisco representa multiplicação);
- **Ajuste Quadrático:** linha verde, ajuste quadrático (grau-2), do tipo y = a0 + a1*x + a2*x^2 (função quadrática genérica, poderia ser analogamente escrita como y = C + Bx + Ax2, onde os termos que acompanham o x são parâmetros a serem acertados ao longo do processo de ajuste);
- **Ajuste Cúbico:** linha roxa, ajuste cúbico (grau-3), do tipo y = a0 + a1*x + a2*x^2 + a3*x^3 (função quadrática genérica).

Tais ajustes no SciDAVis são obtidos a partir do gráfico selecionado e navegando pelos menus *Analysis, Quick Fit, Fit Linear* para ajuste linear; por sua vez se utiliza *Quick Fit, Polynomial Fit*, escolhendo a ordem-2 e a ordem-3 para os ajustes quadráticos e cúbicos. Ainda sobre os ajustes do gráfico é importante notar que os ajustes linear e quadrático são visualmente próximos e não cortam a maioria dos pontos, o que sugere uma correlação próxima entre tais ajustes para o presente exemplo e menor que o ajuste cúbico.

Sobre ajustes **polinomiais**, vale ressaltar!

DESTAQUE

Quanto maior o grau do ajuste, maior é a possibilidade de correlação, visto que polinômios são funções sinuosas e que essa qualidade é determinada pelo grau. Todavia, também é necessário observar que quanto mais elevado o grau de um ajuste obtido, mais difícil se torna a análise humana deste resultado, fato que também pode ser identificado pelo aumento de parâmetros com o grau, ou conceitualmente pelo fato de que uma função não linear varia muito e essa variação não é constante.

Você chegou ao fim desta unidade. Parabéns! Acesse a atividade avaliativa no ambiente virtual de aprendizagem para refletir sobre o que foi apresentado até aqui.

Referências

ALVES, Isabel Fraga. Data Science, Big Data e um novo olhar sobre a Estatística. **Boletim SPE**: O Tema Central da Estatística - um novo olhar, Lisboa, v. 12, n. 2, p. 29-31, 2017. Semestral.

CARVALHO, Marilia Sá; SOUZA-SANTOS, Reinaldo. Análise de dados espaciais em saúde pública: métodos, problemas, perspectivas. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 21, p. 361-378, 2005.

COMARELA, Giovanni et al. Introdução à Ciência de Dados: Uma Visão Pragmática utilizando Python, Aplicações e Oportunidades em Redes de Computadores. SCHAEFFER FILHO, Alberto Egon; CORDEIRO, Weverton Luis da Costa; CAMPISTA, Miguel Elias Mitre (ed.). **Minicursos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**, Porto Alegre, Sociedade Brasileira de Computação, p. 246-295, 2019.

GIBBS, Graham. **Análise de dados qualitativos**. Porto Alegre, Artmed; 2009.

HURWITZ, Judith et al. **Big Data para leigos**. Rio de Janeiro, Alta Books Editora, 2016.

INSTITUTO BRASILEIRO DE GEOGRAFIA (IBGE). **O que é o PIB**. Rio de Janeiro, IBGE, 2021. Disponível em: https://www.ibge.gov.br/explica/pib.php. Acesso em: 12 nov. 2021.

LEITE, Angela; FUITA, Hiroco. **Aplicações da matemática: administração, economia e ciências contábeis**. Boston, Cengage Learning, 2008.

PINTO, José Carlos; SCHWAAB, Marcio. **Análise de Dados Experimentais: I. Fundamentos de Estatística e Estimação de Parâmetros**. Rio de Janeiro, Editora E-papers, 2007.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business**: What you need to know about data mining and data-analytic thinking. Sebastopol (USA) O'Reilly Media, Inc., 2013.

SILVESTRE, António. **Análise de dados e estatística descritiva**. Forte da Casa, Escolar Editora, 2007.

SOUZA, Emanuel Fernando Maia de; PETERNELLI, Luiz Alexandre; MELLO, Márcio Pupin de. Software Livre R: aplicação estatística. 2014. Universidade Federal da Paraíba.

VITALI, Marieli Mezari. Estatística sem matemática para psicologia. **Revista Brasileira de Psicodrama**, São Paulo, v. 27, n. 1, p. 139-144, 2019.

VUOLO, José Henrique. **Fundamentos da teoria de erros**. São Paulo, Editora Blucher, 1996.