ESTATÍSTICA APLICADA AO DATA SCIENCE

PREDIÇÕES COM MODELOS DE REGRESSÃO LINEAR

Autor: Ph.D. Antonio Gomes de Mattos Neto

Revisor: Antonio do Nascimento Alves

INICIAR

introdução

Introdução

Nesta unidade, aprenderemos as “Predições com Modelos de Regressão Linear”. Modelos de regressão linear são um dos principais métodos preditivos da estatística e da ciência dos dados (*data science*). São muito usados em praticamente todos os campos de conhecimento humano (saúde, engenharia, economia, geologia etc.), onde se quer estimar o valor de uma variável quantitativa em função de outras variáveis, chamadas de variáveis de entrada ou preditoras.

Veremos um pouco sobre a evolução recente da estatística e como a mesma se tornou uma das principais ferramentas da ciência dos dados. Falaremos sobre a fusão de métodos de *machine learning* com aqueles da estatística, todos esses usados na ciência dos dados.

Feita essa passagem introdutória por parte do mundo da estatística, de machine learning e ciência dos dados, voltaremos nossas atenções ao principal foco dessa unidade, que são os modelos de regressão linear. Apresentaremos o que são esses modelos, em que situações são usados, tudo isso ilustrado com um exemplo que nos acompanhará ao longo desta unidade. Praticaremos esse aprendizado com atividades a serem realizadas por você.

Estatística, *Machine Learning* e Ciência dos Dados

Nesta seção, discutiremos sobre a relação entre a estatística, a ciência da computação e a ciência dos dados. Falaremos sobre como a estatística e a ciência dos dados fizeram proveito dos algoritmos de aprendizagem de máquina (*machine learning*) da ciência da computação.

Breve Histórico

O nome dessa disciplina é “Estatística Aplicada ao Data Science”. Temos, aqui, a fusão de duas áreas, a estatística e a ciência dos dados, a primeira aplicada à segunda. A estatística é uma área de conhecimento humano mais antigo que a ciência dos dados. Sua estrutura atual começou a tomar forma há aproximadamente 130 anos. Verdadeiramente, sua origem se estende por muitos séculos atrás, mas foram os trabalhos de grandes nomes, tais como os famosos Karl Pearson e Ronald Fisher, que começaram a dar, à estatística, a forma como a conhecemos hoje. Uma ciência forte, com brilho próprio e enorme relevância para a sociedade humana.

Uma outra ciência de enorme relevância para a sociedade moderna é a ciência da computação. Essa é mais recente que a estatística, e só surge com o advento dos computadores, a partir das décadas de 1940 e 1950. A ciência da computação preocupou-se, inicialmente, com temas ligados à arquitetura e funcionamento dos computadores. Porém, em algum momento, passou a tentar emular a inteligência humana. Essas tentativas levaram ao nascimento de uma nova área de conhecimento humano denominada de inteligência artificial, uma área muito vasta e diversificada. Dentro dela, como um dos seus ramos, surgiram os algoritmos de aprendizado de máquina, em inglês *machine learning*.

Mas por que falamos aqui de *machine learning*? Porque entre ciências não há fronteiras rígidas, e tanto os cientistas da computação tomaram emprestados os modelos já desenvolvidos pela estatística, quanto os estatísticos tomaram emprestados os algoritmos de *machine learning* desenvolvidos pela ciência da computação. Essa fusão mostrou-se ser muito rica, vigorosa, e pavimentou o caminho para o surgimento de uma nova área chamada de ciência dos dados.

A ciência dos dados é muito recente na história da sociedade humana. De fato, parece ter surgido entre 20 a 10 anos atrás. Um dos relatos que se ouve é que uma das grandes empresas americanas da era digital anunciou seu interesse em contratar “cientistas de dados”. Mas quem ela queria contratar? Bem, parece que ela queria contratar um estatístico, mas um estatístico com um viés computacional forte. Um que soubesse programar. Mas ela também ficaria satisfeita com um cientista da computação, porém um que possuísse algum conhecimento de estatística, pois um cientista da computação sabe programar muito bem, mas para se tornar um cientista de dados precisa conhecer estatística.

É por esse motivo que a estatística está intimamente ligada à ciência dos dados e ao mundo dos algoritmos de *machine learning* da ciência da computação. Enfim, todos, de alguma forma, entrelaçados. Cada um desses mundos com suas especialidades, mas usufruindo mutuamente dos conhecimentos gerados pelos outros três mundos: 1) a estatística, com seus métodos tão cuidadosamente construídos e aplicados; 2) a ciência da computação, com seus algoritmos de *machine learning*; e 3) a ciência dos

dados, que aplica todos esses conhecimentos e métodos de forma fértil e produtiva.

Linguagens de Programação na Ciência dos Dados

Conversamos cada vez mais com as máquinas (sejam computadores, *smartphones*, nossos carros etc.), e as máquinas entre si (IoT = *Internet of Things*, a Internet das Coisas). Essa conversa com as máquinas e a conversa delas entre elas mesmas é construída por meio de linguagens de programação que se transformam em códigos executáveis e permitem a realização das comunicações dos seres humanos com as máquinas e das máquinas entre si.

São tão diversas as linguagens de programação, e suas histórias tão variadas, que seria impossível tentarmos abordar esse assunto aqui. Mas aqui devemos deixar claro que não se faz mais estatística sem o uso intensivo de computadores e, para isso, precisamos lançar mão de linguagens de programação. E isso também se aplica à ciência dos dados. Sendo assim, que linguagens de programação são as mais empregadas por essas duas ciências, a ciência dos dados e a estatística?

A resposta é: Para rodar poderosas rotinas computacionais usam-se linguagens como Fortran, C, C++ e Java, e para o desenvolvimento de aplicações em ciência dos dados, linguagens mais flexíveis como R, Python, Julia e MatLab são preferidas.

Dentre essas, R e Python são aquelas que se destacam no mundo da estatística e da ciência dos dados. Ambas incríveis e muito produtivas. Aqui, nesta disciplina, será mais fácil usarmos o R. A razão é simples. O R base já vem com todas as funções estatísticas e gráficas das quais precisaremos. Mas não se preocupe, usaremos o R apenas para alguns exercícios bem simples, só para mostrar a você que é possível. Além disso, é muito fácil instalar e rodar o R.

Em suma, preste muita atenção a essas duas linguagens de programação: o R e o Python. O mercado valoriza quem possui alguma familiaridade com elas e com a estatística e a ciência dos dados.

praticar

Vamos Praticar

Campo em crescimento exponencial, a Ciência dos Dados tem se tornado uma área apaixonante para entusiastas das mais diversas áreas. Estatísticos contribuem com sólida teoria de análise de dados enquanto cientistas e engenheiros da computação contribuem com novas capacidades e possibilidades computacionais. Assim, pesquisadores(as) da biologia, psicologia, direito, economia, comunicação, sociologia e diversas outras áreas podem usufruir desse conjunto de técnicas (algumas nem tão novas) para aprimorar e desenvolver suas pesquisas. E a linguagem R é uma das principais linguagens de programação utilizadas para isso.

STABILE, M. Prefácio. *In*: OLIVEIRA, P. F. de; GUERRA, S.; MCDONNEL, R. **Ciência dos Dados com R**: introdução. Brasília: IBPAD, 2018. p. 7.

Diante disso, assinale a alternativa correta:

- ○ **a)** A ciência dos dados desenvolveu-se sem qualquer ligação com a estatística ou com a ciência da computação.
- ○ **b)** Linguagens de programação não têm relevância nas aplicações da estatística à ciência dos dados.
- ○ **c)** O R não é uma linguagem de programação de referência para a estatística e a ciência dos dados.
- ○ **d)** A ciência dos dados, campo que vem apresentando um crescimento exponencial, pode ser usufruída por pesquisadores e profissionais das mais diversas áreas.
- ○ **e)** Não é possível a aplicação da estatística ou da ciência dos dados em ciências humanas ou sociais, mas apenas nas ciências exatas.

Predição com Regressão Linear - Estudo de Caso

Nesta seção, discutiremos sobre o desenvolvimento de modelos preditivos e apresentaremos o estudo de caso que nos acompanhará ao longo desta unidade. Através desse caso, aprenderemos as regressões linear e múltipla.

Fases do Desenvolvimento de Modelos Preditivos

Nesta seção, veremos como modelos de regressão linear podem nos ajudar a predizer o valor de venda de um imóvel, a partir de dados coletados relativos a algumas de suas características. Faremos isso por meio de um exemplo ilustrativo, centrado em dois personagens principais: uma corretora de imóveis, especializada na venda de apartamentos, e um estatístico. Esse exemplo, um estudo de caso simulado, nos acompanhará até o final da unidade. Ele vai nos permitir entender sobre uma das maneiras como a estatística pode ser aplicada à ciência dos dados.

Na criação de uma aplicação que tem como objetivo o desenvolvimento de um algoritmo preditivo, uma das maneiras possíveis de descrever as

principais etapas do seu desenvolvimento é:

1. Definição da questão a ser resolvida (*business case*);
2. Definição dos dados necessários ao desenvolvimento do caso;
3. Coleta dos dados (evitando vícios de amostragem);
4. Limpeza e tratamento dos dados (quando necessário);
5. Análise descritiva (resumos estatísticos e visualização gráfica);
6. Escolha de um modelo (algoritmo) preditivo;
7. Ajuste (treino) do modelo (do algoritmo);
8. Teste do modelo para verificação da sua *performance* preditiva;
9. Entrega do modelo para validação e utilização pelo cliente;
10. *Feedback* do cliente para ajustes e melhoramentos do modelo.

Como, neste material, pretendemos dar apenas uma visão inicial, e também como não podemos nos estender com mais profundidade em aspectos mais técnicos, não daremos atenção a todas essas etapas. Concentraremo-nos em mostrar, de uma forma mais direta e simples possível, o poder de predição de modelos de regressão linear, em situações típicas onde podem ser empregados. Se você tiver interesse em se aprofundar nesses temas, há uma vasta literatura disponível para sua consulta e leitura como, por exemplo, o livro Ciência dos Dados - Introdução (OLIVEIRA; GUERRA; MCDONNEL, 2018).

Predição do Valor de Venda de Imóveis

Uma corretora de imóveis residenciais queria saber se seria possível, através de algum tipo de aplicativo, estimar valores de venda de imóveis residenciais. Ela trabalhava exclusivamente com venda de apartamentos e sabia que, se pudesse contar com tal aplicativo, teria mais agilidade na definição do valor de venda dos imóveis junto aos proprietários, assim como tornaria mais assertiva sua conversa junto a compradores potenciais. Em outras palavras, imaginava que com tal aplicativo ela teria mais e melhores argumentos, pois seriam argumentos balizados tecnicamente. Sendo mais precisa, imaginava que poderia dar mais velocidade aos negócios, gerando mais satisfação para os seus clientes, assim como melhores resultados para a imobiliária onde trabalhava já há anos.

Como tinha um amigo estatístico, decidiu conversar com ele. Seu amigo estatístico lhe pediu para trazer alguns dados de mercado. Passados alguns dias, a corretora voltou a procurar seu amigo estatístico e lhe mostrou os seguintes dados, observados de 100 imóveis residenciais – todos apartamentos, sua especialidade – vendidos nos últimos meses pela imobiliária. No Quadro 1.1 exibimos algumas do total das 100 observações que ela coletou:

Ap.	Área (m2)	Andar	Local	Valor (R$mil)
1	59,4	2	Bairro	398
2	62,7	8	Bairro	340
3	80,6	4	Centro	544
4	65,7	9	Bairro	283
...	...	...	...	...
99	62,6	4	Centro	304
100	54,7	6	Centro	347

Quadro 1.1 - Dados amostrados relativos aos apartamentos
Fonte: Elaborado pelo autor.

Com isso, o estatístico tinha, em suas mãos, dados. O estatístico podia, a partir desse momento, examinar esses dados e decidir sobre o que fazer, o que veremos em seguida, após uma atividade para você treinar seus conhecimentos.

praticar

Vamos Praticar

Workflow da Ciência dos Dados: não existe apenas uma forma de estruturar e aplicar os conhecimentos da Ciência dos Dados. A forma de aplicação varia bastante conforme a necessidade do projeto ou do objetivo que se busca alcançar. Neste curso, usaremos um modelo de workflow bastante utilizado. Esse workflow propõe basicamente os seguintes passos: Carregar os dados; Limpar os dados; Transformar, visualizar e modelar; Comunicar o resultado.

OLIVEIRA, P. F. de; GUERRA, S.; MCDONNEL, R. **Ciência dos Dados com R**: introdução. Brasília: IBPAD, 2018. p. 10.

Descrevemos as principais etapas para o desenvolvimento de um algoritmo preditivo na ciência dos dados. Deixamos claro que as etapas lá descritas são apenas uma das formas de se definir essas etapas, porém, entre todas as descrições, há similaridades. Veja, por exemplo, a descrição dessas etapas como dadas no texto introdutório referenciado. Analise as duas descrições (do *e-book* e do texto introdutório), reflita e assinale a alternativa correta:

- ○ **a)** No desenvolvimento de um projeto em ciência dos dados nunca se faz a visualização dos dados, também chamada de análise descritiva dos dados.
- ○ **b)** No desenvolvimento de um projeto em ciência dos dados nunca se faz a etapa de limpeza e tratamento dos dados.
- ○ **c)** No desenvolvimento de um projeto em ciência dos dados é comum que se faça a modelagem dos dados, que consiste na escolha, treino e teste de um modelo.
- ○ **d)** No desenvolvimento de um projeto em ciência dos dados nunca se faz a comunicação dos resultados, pois eles só interessam ao próprio cientista de dados.
- ○ **e)** Nas etapas de desenvolvimento de um projeto em ciência dos dados descritas no *e-book* não se considera a etapa de *feedback* do cliente para ajustes e melhoramentos do modelo.

Análise Descritiva dos Dados

Nesta seção, discutiremos sobre a estrutura de dados preferida pelos estatísticos e cientistas de dados e, em seguida, veremos um exemplo de como um cientista de dados examina seus dados através de técnicas descritivas, que são sumários estatísticos, também chamados de resumos, e gráficos para a visualização dos dados.

Dados Retangulares e Data-frames

O estatístico decidiu examinar os dados que sua amiga corretora lhe trouxe. Ele logo percebeu que estavam bem organizados, na forma de uma tabela, com as variáveis dispostas em colunas, e as observações relativas a cada imóvel dispostas em linhas. Essa é, talvez, a forma preferida por um estatístico, ou por um cientista de dados, de organização de dados.

Por vezes, referimo-nos a dados que podem ser organizados em uma tabela desse jeito, ou seja, as variáveis dispostas nas colunas e as observações dispostas nas linhas, como dados retangulares, ou dados estruturados. No *software* estatístico R, essa forma de organização de dados é referida como

"data-frame". Esse conceito foi copiado, alguns anos depois (em 2012), pelo Python, por meio de sua famosa biblioteca "Pandas".

Voltando ao estatístico, ele também viu que sua amostra tinha tamanho $n = 100$, ou seja, lá havia dados relativos a 100 diferentes apartamentos. Para simplificar a sua análise, ele decidiu adotar uma notação compacta para as variáveis observadas:

$$X_1 = \text{área do imóvel } (m^2)$$

$$X_2 = \text{andar do imóvel } (1, 2, 3, \ldots)$$

$$X_3 = \text{localizaão do imóvel (Bairro ou Centro)}$$

$$Y = \text{valor de venda do imóvel (kR\$)}$$

Dados Relativos à Área do Imóvel

O estatístico iniciou sua análise examinando X_1 (área do imóvel). Usou as funções min(), mean(), max() e sd() do software estatístico R na determinação dos valores mínimo, médio, máximo e desvio-padrão dos dados observados para x_1:

$$\text{min}(x1) = 41,9 \quad \text{mean}(x1) = 65,6 \quad \text{max}(x1) = 86,9 \quad \text{sd} = 9,1$$

Ele viu, então, que para esses 100 apartamentos vendidos, a área variou entre um mínimo de $41,9\ m^2$ e um máximo de $86,9\ m^2$, com área média de $65,6\ m^2$ e desvio-padrão da área de $9,1\ m^2$.

Em seguida, decidiu visualizar esses dados. Como área é uma variável quantitativa, optou por construir um histograma de X_1 usando a função gráfica hist() do R:

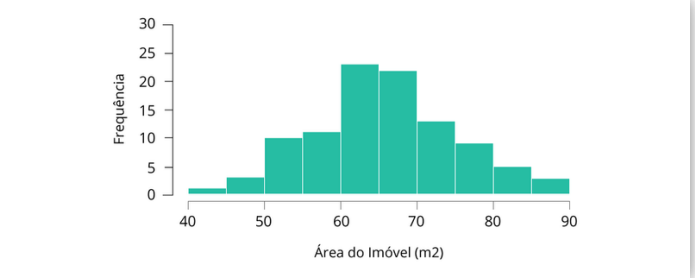


Figura 1.1 - Histograma dos dados relativos à área dos apartamentos
Fonte: Elaborada pelo autor.

Dados Relativos ao Andar do Imóvel

Depois o estatístico passou ao exame dos dados relativos à variável X_2 (andar do imóvel). Também aqui usou de algumas funções do R para calcular o valor mínimo (min), a mediana (median), o máximo (max) e o desvio-padrão (sd) dos dados observados:

$$\text{min}(x2) = 1 \quad \text{median}(x2) = 4 \quad \text{max}(x2) = 14 \quad \text{sd} = 3,1$$

Ele viu, então, que para esses 100 apartamentos vendidos, o andar do imóvel variou entre um mínimo de 1 (primeiro andar), uma mediana de 4 (50% dos apartamentos até o quarto andar), um máximo de 14 (décimo-quarto andar) e um desvio-padrão de 3,1 andares (uma indicação da variabilidade desses dados relativos ao andar dos apartamentos).

Em seguida decidiu visualizar esses dados. Como o andar do imóvel é uma variável quantitativa, optou por construir um histograma de X_2 usando a função gráfica hist() do R:

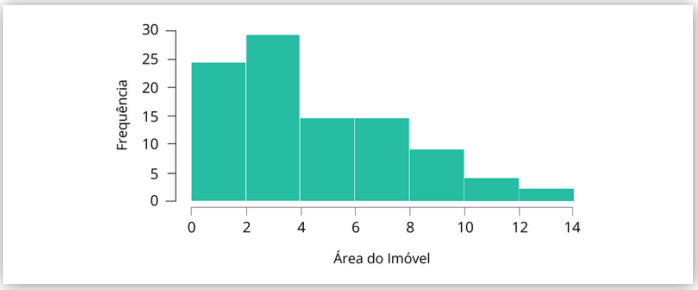


Figura 1.2 - Histograma dos dados relativos ao andar dos apartamentos
Fonte: Elaborada pelo autor.

Dados Relativos à Localização do Imóvel

Em seguida, o estatístico passou ao exame dos dados relativos à localização dos apartamentos. Logo percebeu que essa variável, X_3 (local do imóvel) tratava-se de uma variável qualitativa nominal com apenas dois níveis,

“Bairro” e “Centro”, uma variável por vezes chamada de dicotômica, em oposição às variáveis qualitativas politômicas, as quais podem assumir mais de dois níveis, ou classes. Decidiu codificar esses dados usando uma forma de codificação muito comum para variáveis dicotômicas, como segue:

Bairro = 0 Centro = 1

Após isso, como esses dados são qualitativos, uma das formas mais práticas para sumarizá-los é contando a frequência de aparição de cada nível (0 ou 1) na amostra coletada. Para isso usou de uma interessante função do R, denominada de table(), obtendo os seguintes resultados:

table(x3) 0 1

32 68

Ou seja, do total de apartamentos observados, 32 estavam localizados no Bairro (0) e 68 no Centro (1). A corretora havia explicado, ao estatístico, que a imobiliária havia coletado os dados dessa forma, sem tentar distinguir em maior detalhe qual exato bairro ou qual exato local no centro, porque com base na sua experiência de vários anos, havia concluído não haver a necessidade de maior detalhamento, ao menos naquele município onde ela atuava.

Em seguida decidiu visualizar esses dados. A forma preferida do estatístico ou do cientista de dados de visualizar dados qualitativos é por meio de diagramas de barras. Nesse diagrama, cada nível (classe) da variável é associada a uma barra, e a altura da barra é proporcional à frequência absoluta com que o nível (classe) foi observado na amostra.

O estatístico usou uma função gráfica do R, denominada de barplot(), e obteve o seguinte resultado:

Como você pode observar, o diagrama de barras oferece uma simples, porém bastante efetiva, visualização da frequência de observações de cada nível (classe) da variável qualitativa. Vale notar aqui que podemos usar a frequência relativa no lugar da frequência absoluta, com o mesmo resultado visual. Também vale notar que gráficos de pizza são uma alternativa aos diagramas de barras.

Dados Relativos ao Valor de Venda do Imóvel

Finalmente, o estatístico prosseguiu com sua análise descritiva examinando a variável Y (valor de venda do imóvel) e, novamente, usou de funções do R para calcular os valores mínimo (min), médio (mean), máximo (max) e desvio-padrão (sd) dos dados observados:

$$\text{min}(y) = 129 \quad \text{mean}(y) = 366,5 \quad \text{max}(y) = 556 \quad \text{sd} = 85,9$$

Vemos que, para esses 100 apartamentos, o valor de venda variou entre um mínimo de 129 kR$ e um máximo de 556 kR$, com valor médio de 366,5 kR$ e um desvio padrão de 85,9 kR$.

Assim como fez para as outras variáveis, também aqui resolveu visualizar os dados coletados quanto ao valor de venda. Sendo esses dados quantitativos, construiu um histograma de Y (valor de venda do imóvel) usando a função gráfica hist() do R:

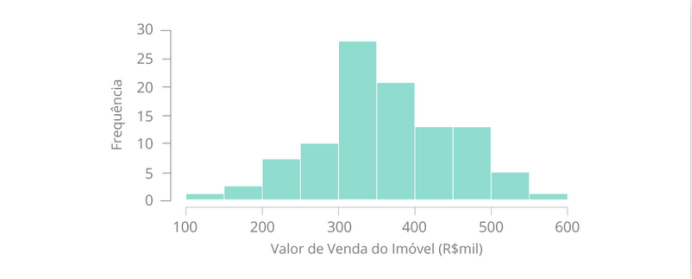


Figura 1.4 - Histograma dos dados do valor de venda dos apartamentos
Fonte: Elaborada pelo autor.

Percebeu serem dos dados relativos aos valores de venda dos imóveis distribuídos de forma ligeiramente assimétrica, com uma maior frequência de observações se concentrando à direita.

Visualização do Valor ***versus*** Área do Imóvel

Como a ideia da corretora era conseguir fazer uma predição do valor de venda de um apartamento dadas as suas características (com base nos dados coletados na amostra), o estatístico decidiu visualizar essa possível relação construindo um gráfico de dispersão (*scatter plot em inglês*), no qual plotaria os dados relativos à área do imóvel X_1 no eixo horizontal, e os dados relativos ao valor de venda do imóvel Y no eixo vertical. Gráficos de dispersão são usados para a visualização da relação entre variáveis quantitativas. Os dados, nesse caso, devem ser tomados aos pares, isto é (X_1, Y), a primeira e a última coluna da tabela, linha a linha (aos pares):

Quadro 1.2 - Tabela área e valor
Fonte: Elaborado pelo autor.

Para isso o estatístico usou uma função gráfica do R de nominada de plot(), obtendo o seguinte resultado exibido na Figura 1.5. Essa figura mostra que há uma associação positiva entre Y e X_1. Há uma tendência de Y subir (o valor do imóvel), quando X_1 cresce (a área do imóvel). A dispersão dos dados se dá porque há outros fatores influentes que causam essa variabilidade nas observações da amostra coletada. A função cor() do R permite uma medida da força dessa associação:

$$\text{cor}(y, x1) = 0,55$$

Esse valor indica que, para os dados amostrados, a correlação entre Y e X_1 é positiva, com uma força moderada.

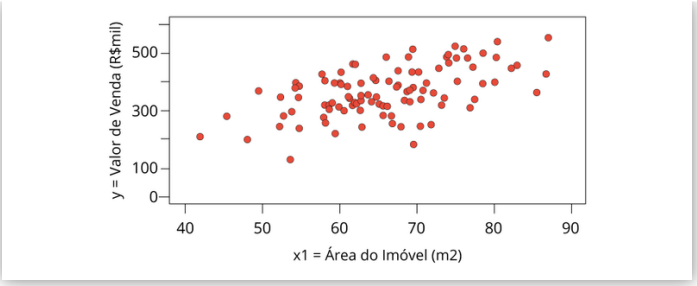


Figura 1.5 - Gráfico de dispersão da área e valor dos apartamentos
Fonte: Elaborada pelo autor.

Visualização do Valor ***versus*** o Andar do Imóvel

Em seguida, fez o mesmo para a relação entre os dados de valor de venda do imóvel Y versus seu andar X_2, e os plotou aos pares, isto é, (X_2, Y), a segunda e a última coluna da tabela com os dados dos imóveis, linha a linha (aos pares). Novamente o estatístico recorreu à função plot() do R e obteve o resultado exibido na Figura 1.6.

Essa figura também mostra que há uma associação positiva entre Y e X_2, porém mais sutil. Talvez você não consiga ver isso muito bem, mas não se preocupe. O estatístico tem já uma grande experiência, e mesmo ele pode ter dificuldade em ver que há uma tendência de Y subir (o valor do apartamento), quando X_2 cresce (o andar do apartamento). Para verificar essa questão, o estatístico aqui lançou mão da função cor() do R, obtendo:

$$\text{cor}(y, x2) = 0,24$$

Também nesse caso a correlação é positiva, porém aqui com uma força mais fraca do que a correlação entre Y e X_1. A dispersão dos dados se dá porque há outros fatores influentes e ruídos, que causam variabilidade nas observações da amostra coletada.

Outra curiosidade é que os dados relativos à Y encontram-se “empilhados” sobre alguns valores de X_2, mas isto é apenas fruto direto do fato que X_2 varia de forma discreta, ou seja, $X_2 = 1, 2, 3, \ldots$, o andar de cada apartamento vendido.

Visualização do Valor ***versus*** Localização

Aqui o estatístico teve de lançar mão de um tipo de gráfico que permitisse a visualização de dados quantitativos Y (valor de venda) versus dados qualitativos X_3 (localização do imóvel). Uma solução muito inteligente para isso é recorrer aos boxplots (diagramas de caixas), onde no eixo horizontal indicamos os níveis da variável qualitativa X_3 e no eixo vertical os valores observados da variável quantitativa Y, também aos pares (X_3, Y), isto é, a terceira e a quarta coluna da tabela de dados. O resultado que o estatístico obteve foi o seguinte:

Esse gráfico permite ver como se dispersam os valores de venda dos imóveis da amostra, exibidos ao longo do eixo vertical, em função da sua localização, exibida no eixo horizontal. Veja que imóveis no centro têm valor inferior a imóveis no bairro. Nas palavras de um especialista: “Boxplots são muito úteis na visualização gráfica entre diferentes conjuntos de dados, porque têm um alto impacto visual e são fáceis de entender” (MONTGOMERY, 2013, p. 139). São muito usados nas situações em que queremos visualizar a relação de dados quantitativos com dados qualitativos.

praticar

Vamos Praticar

Dados Estruturados: talvez seja o formato mais fácil de se trabalhar no R. São conjuntos de informações organizadas em colunas (atributos, variáveis, *features* etc.) e linhas (registros, itens, observações etc.). São dados mais comumente encontrados diretamente em bancos de dados, arquivos com algum tipo de separação entre as colunas, Excel, arquivos com campos de tamanhos fixo etc.

OLIVEIRA, P. F. de; GUERRA, S.; MCDONNEL, R. **Ciência dos Dados com R**: introdução. Brasília: IBPAD, 2018. p. 19.

Com base neste texto, assinale a alternativa correta:

- ○ **a)** Dados estruturados não são importantes para a estatística ou para a ciência dos dados.
- ○ **b)** No R, estruturas de dados organizados em tabelas, com as variáveis dispostas nas colunas e as observações nas linhas, são chamados de data-frames.
- ○ **c)** Dados retangulares não são a mesma coisa que dados estruturados. Dados retangulares têm estrutura de retângulos, e dados estruturados são organizados na forma de tabelas.
- ○ **d)** Dados retangulares não são organizados com as variáveis dispostas nas colunas e as observações dispostas nas linhas.

○ **e)** Chamar variáveis de atributos é impróprio na ciência dos dados, já que variáveis são sempre variáveis.

Predição com Modelos de Regressão Linear

Nesta seção, veremos como modelos de regressão linear simples e múltipla são empregados como modelos preditivos de valores de variáveis quantitativas. Isso será ilustrado com o desenvolvimento de um modelo de regressão linear na predição do valor de venda de imóveis.

Modelos de Regressão Linear

Modelos de regressão linear são usados para a predição do valor esperado de uma variável resposta quantitativa, habitualmente anotada como Y, em função de uma ou muitas variáveis de entrada, habitualmente anotadas como X, com um índice a elas associados se mais do que uma. Por exemplo, no caso aqui estudado, temos três variáveis de entrada, X_1, X_2 e X_3. Esquematicamente, podemos representar essa ideia da seguinte forma:

Figura 1.8 - Representação da transformação das entradas na saída
Fonte: Elaborada pelo autor.

O modelo aqui funciona como uma função que transforma os dados de entrada em um dado de saída. Vale dizer, nesse momento, que há outras denominações comuns para essas variáveis, tais como:

$$X = \text{variável de entrada, regressora, preditora, independente}$$

$$Y = \text{variável de saída, de resposta, dependente, target variable}$$

Vamos ver como evolui o caso da nossa corretora, seus apartamentos e a ajuda do seu amigo estatístico.

Valor ***versus*** Área do Imóvel

Em um primeiro momento, a corretora pediu para o estatístico fazer uma tentativa inicial de predição usando apenas a variável área do imóvel X_1 como variável de entrada (preditora).

O estatístico imediatamente pensou em um modelo de regressão linear simples. O nome simples, na regressão linear, significa que o modelo de regressão considerará apenas uma variável de entrada (aqui, neste nosso caso, X_1, a área do imóvel) e procurará verificar qual seu possível efeito na variável resposta (aqui, neste nosso caso, Y, o valor do imóvel), com base nos dados amostrados.

O estatístico, então, escreveu o seguinte modelo de regressão linear simples para essa situação:

$$y = b_0 + b_1 x_1$$

Aqui, b_0 e b_1 são coeficientes do modelo. Seu maior interesse, nesse momento, era o de determinar os valores desses coeficientes. Com isso, ele poderia estimar $y = E[Y]$, o valor esperado (valor médio) para o imóvel, quando sua área X_1 for igual a x_1 metros quadrados, ou seja, $X_1 = x_1$.

Ele fez isso usando de um método clássico da estatística, o Método dos Mínimos Quadrados. Não é nosso objetivo discutir o funcionamento desse método, mas apenas ilustrar o poder da estatística quando aplicada à ciência dos dados. Vamos nos concentrar nos resultados da aplicação desse método quando o usamos para o cálculo dos coeficientes b_0 e b_1. O estatístico usou do software R para fazer esses cálculos, e obteve:

$$b0 = kR\$27,22 \quad \text{e} \quad b1 = kR\$5,15/m2$$

tal que, substituindo esses valores no modelo de regressão linear simples acima, chegamos a:

$$y = 27,22 + 5,15\,x_1$$

Esse resultado pode ser plotado no gráfico de dispersão que vimos anteriormente para o valor do imóvel y, em função da área do imóvel x_1:

Vemos que a plotagem do modelo ajustado fornece uma reta, com interseção com o eixo vertical em $x_1 = 0$ igual a $b_0 = 27,22\,m^2$ e inclinação igual a $b_1 = kR\$\,5,15/m^2$. Podemos mudar a escala do eixo horizontal para a mesma

escala que usamos anteriormente para a construção do gráfico de dispersão entre y e $x1$, resultando numa melhor visualização:

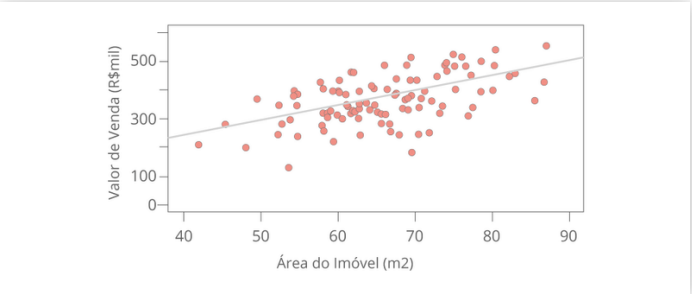


Figura 1.10 - Gráfico de dispersão da área e valor dos apartamentos.
Fonte: Elaborada pelo autor.

Devemos interpretar esse resultado. O coeficiente de interseção é o valor esperado (valor médio) para y quando $x1 = 0$, ou seja, o valor esperado para o preço de venda quando a área do apartamento for igual a zero. Essa interpretação não tem um sentido real, pois não existem apartamentos com área igual a zero. Nessa situação, é comum tomarmos esse coeficiente apenas como um coeficiente de ajuste do modelo, sem nos preocuparmos em atribuir a ele um significado “real”.

Apenas quando faz sentido a variável preditora assumir um valor igual a zero, é que também faz sentido interpretar o coeficiente b0 não só como um coeficiente de ajuste do modelo, mas efetivamente como o valor esperado para y quando $x1$ é igual a zero. Espero que você tenha entendido esse ponto. Não é muito complicado.

Já com respeito ao coeficiente $b1$, que é a inclinação da reta, esse sempre terá uma interpretação bastante útil. O valor calculado pelo estatístico para esse coeficiente foi:

$$b1 = kR\$5,15/m2$$

Ele representa o quanto aumenta o preço de venda do apartamento com o aumento da área em exatamente 1 metro quadrado. Isto é, ele é o valor do

metro quadrado médio para os apartamentos da amostra que a corretora passou para o estatístico.

Fazer uma predição do valor esperado de y (preço de venda do imóvel), dada sua área em x_1 em metros quadrados, fica fácil agora. Suponha que você quer saber qual seria o preço de venda médio estimado para um apartamento de $65\ m^2$. Basta substituir esse valor na equação do modelo e o resultado será

$$y = 27,22 + 5,15 \times 65 = 362$$

Aqui, arredondamos o valor 361,97 mil para 362 mil reais, pois estamos estimando em mil reais, e não temos interesse em frações de mil reais.

Valor ***versus*** Andar do Imóvel

O estatístico mostrou à sua amiga corretora a análise preditiva que ele havia realizado com base nos dados da área dos imóveis x_1 e seus valores de venda y. Ela ficou muito admirada e curiosa em saber como seria esse resultado se, ao invés de usarmos como dados de entrada a área dos imóveis, usássemos o número x_2 do seu andar. E pediu que o estatístico desenvolvesse esse outro modelo preditivo.

Obviamente, o estatístico, já tendo usado um modelo de regressão linear simples para a situação anterior, decidiu fazer o mesmo para esse novo caso, e escreveu o seguinte modelo de regressão linear simples para essa nova situação:

$$y = b_0 + b_2 x_2$$

Também aqui aplicou o Método dos Mínimos Quadrados para o ajuste do modelo, por meio do software estatístico R. Obteve os seguintes valores para os coeficientes do modelo:

$$b_0 = kR6,55/andar$$

tal que, substituindo esses valores no modelo de regressão linear simples acima, chegamos a:

$$y = 333,71 + 6,55\, x_2$$

Esse resultado pode ser plotado no gráfico de dispersão que vimos anteriormente para valor do imóvel y (kR$) em função de andar do imóvel x_2 (1, 2, 3, ...):

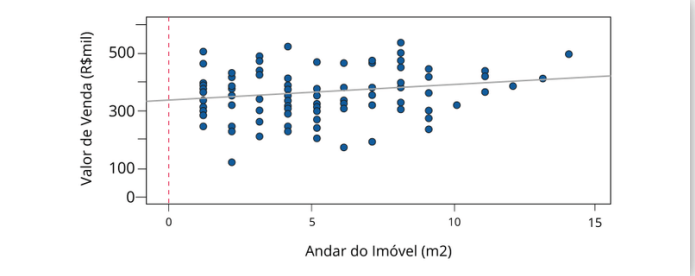


Figura 1.11 - Gráfico de dispersão do andar e valor dos apartamentos
Fonte: Elaborada pelo autor.

Também aqui devemos interpretar esse resultado. O coeficiente de interseção é o valor esperado para y quando $x_2 = 0$, ou seja, o valor esperado para o preço de venda quando o andar do apartamento for igual a zero, o térreo. Nesse caso, temos uma interpretação para o coeficiente de interseção do modelo, além de um mero parâmetro de ajuste do modelo aos dados amostrados, pois existem apartamentos em andares térreos. Devemos notar, entretanto, que na amostra coletada pela imobiliária onde trabalha a corretora, nenhum dos apartamentos vendidos ficava no andar térreo. Fazer $x_2 = 0$, nesse caso, é uma extrapolação da predição para além da região onde os dados foram observados. Quando $x_2 = 0$ (andar térreo), a predição para o valor do imóvel é

$$y = 333,71 + 6,55 \times 0 = 333,71$$

Ou seja, y é exatamente igual a b_0, a interseção da reta com o eixo vertical na posição $x_2 = 0$ do gráfico.

Quanto ao coeficiente b_2, que é a inclinação da reta, a interpretação é similar àquela que já demos anteriormente para o caso do coeficiente b_1. O valor calculado para esse coeficiente foi:

b2 = kR 6,55/andar representa o valor do aumento por cada 1 andar (valor unitário por andar), o qual deve se somar a R$ 333,71 mil para se ter a estimativa do valor esperado para o valor do imóvel.

Fazer uma predição do valor esperado de y (preço de venda do imóvel), dado seu andar, fica fácil agora. Suponha que você quer saber qual seria o preço de venda médio estimado para um apartamento no décimo andar. Basta substituir esse valor na equação do modelo e o resultado será

$$y = 333,71 + 6,55 \times 10 = 399$$

Aqui, arredondamos o valor 399,21 para 399 mil reais, pois estamos estimando em mil reais, e não temos interesse em frações de mil reais.

Valor ***versus*** Área e Andar do Imóvel

Nesse ponto o estatístico decidiu combinar os dois modelos anteriores em um só, onde o valor esperado para y (valor do imóvel) é escrito como função de x_1 (área do imóvel) e x_2 (andar do imóvel), simultaneamente. Esse modelo fica assim:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Denominamos um modelo desse tipo, onde há mais do que uma variável de entrada, de modelo de regressão linear múltipla. Muito importante é evitarmos a tentação de usar os valores previamente determinados, nos modelos de regressão simples anteriores, para b_0, b_1 e b_2 , nesse modelo de regressão múltipla. Quando aplicamos o Método dos Mínimos Quadrados, cada novo modelo deve ser ajustado aos dados da amostra independentemente de outros modelos, gerando assim um conjunto de coeficientes específicos para si.

O estatístico, que conhecia muito bem sobre isso, recorreu novamente ao software estatístico R para calcular os valores dos coeficientes desse novo modelo. Chegou aos seguintes resultados:

$$b0 = -kR5,12/m2b2 = kR\$6,34/andar$$

Substituindo esses valores no modelo de regressão múltipla, temos:

$$y = -2,59 + 5,12\,x1 + 6,34\,x2$$

Essa expressão pode ser usada para fazermos predições do valor esperado de y à área desejada e o andar desejado para o apartamento. A corretora já aproveitou para fazer um teste, pois uma cliente gostaria de saber qual valor esperado de um apartamento com uma área de 50 metros quadrados, situado no 10º andar. Esse apartamento seria para ela, o seu marido e um filhinho. De posse do modelo, foi simples fazer a predição:

$$y = -2,59 + 5,12 \times 50 + 6,34 \times 10 = 317$$

Aqui, novamente, arredondamos 316,81 para 317 mil reais, pois queremos avaliar o valor do imóvel sem nos preocuparmos com frações de mil reais.

Valor ***versus*** Área, Andar e Localização do Imóvel

Naturalmente, a corretora ficou muito feliz ao ver que já dispunha de um algoritmo de predição. Percebeu que agora só faltava incluir no modelo de regressão múltipla a última variável da base de dados da imobiliária, ou seja, aquela relativa à localização do imóvel.

A imobiliária só registrava se o imóvel havia sido vendido em um bairro ou no centro. Sendo assim, essa variável, que é uma variável qualitativa, só podia assumir dois valores (dois níveis, duas classes). Você lembra que o estatístico já havia decidido codificar esses dois níveis da seguinte forma:

Bairro = 0 Centro = 1

O estatístico prosseguiu e escreveu o seguinte modelo de regressão múltipla com variáveis preditoras mistas (quantitativas e qualitativas):

$$y = b0 + b1x1 + b2x2 + b3x3$$

Novamente, recorreu ao R e calculou os coeficientes para esse modelo, obtendo:

b0 = kR 4,87 / m2

b2 = kR 27,43 / localização

Substituindo esses valores na expressão do modelo, fica assim:

$$y = 32,67 + 4,87\,x_1 + 6,36\,x_2 - 27,43\,x_3$$

Já discutimos a interpretação dos coeficientes b_1 e b_2. Vamos, agora, discutir a interpretação do coeficiente b_3. Para isso basta lembrar que x_3 pode assumir dois valores, ou níveis (classes), Bairro = 0 e Centro = 1. Como o coeficiente b_3 está multiplicando x_3, a contribuição do termo contendo b_3 para o valor de y será zero quando $x_3 = 0$ e menos kR$ 27,43 quando $x_3 = 1$. Ou seja, o modelo nos informa que, quando o imóvel está localizado no centro, ele custa, em média, 27,43 mil reais a menos que um apartamento de bairro de mesma área e mesmo andar.

A corretora pediu um exemplo. Precisava entender melhor. O estatístico, então, deu o seguinte exemplo: pensou na mesma cliente que já havia solicitado uma predição do valor esperado para um apartamento de 50 metros quadrados de décimo andar; mas ela não havia especificado onde, se no bairro ou no centro; ora, agora ele tinha um modelo que levava em conta essa variável e só substituiu os 50 metros e 10º andar no modelo. Obteve o seguinte resultado:

$$y = 32,67 + 4,87 \times 50 + 6,36 \times 10 - 27,43\,x_3 = 339,77 - 27,43\,x_3$$

e viu que:

$x_3 = 0$ (bairro) $\Rightarrow$ $y = 339,77$

$x_3 = 1$ (centro) $\Rightarrow$ $y = 312,34$

A diferença de valor é 27,43 mil reais, que resulta em 27 mil reais ao arredondarmos para mil. Apartamentos de mesmas características no centro

custam 27 mil reais a menos que apartamentos nos bairros. Isso vale para aquele município, para os dados amostrados pela imobiliária e para esse modelo específico de regressão linear múltipla, com variáveis de entrada (preditoras) mistas, quantitativas e qualitativas. Outros dados e outros modelos podem levar a resultados diferentes.

A corretora entendeu e quase atingiu o auge de sua felicidade. Agora tinha à sua disposição um algoritmo preditor de valores esperados para os imóveis que ela comercializava. Mas, e o aplicativo?

O aplicativo deve ser produzido em um passo posterior ao desenvolvimento do algoritmo. Com o algoritmo de predição pronto, agora, a corretora deverá procurar um profissional que possa desenvolver um aplicativo (um engenheiro de *software*, por exemplo), especializado em aplicações na web ou em *smartphones*. Esse profissional criará uma interface entre o usuário (a corretora) e o algoritmo (o modelo preditivo), tal que, com a entrada de dados das características de um apartamento, o aplicativo produzirá, na tela do computador ou do *smartphone*, a predição do seu valor esperado (médio) de venda.

reflita

Reflita

Será que você sabia que a estatística e a ciência dos dados são muito usadas nas ciências dos esportes, tanto amadores quanto profissionais? E você? Consegue se imaginar trabalhando para um grande clube como especialista em análise estatística esportiva? Reflita sobre isso, enquanto lê, analisa e pensa sobre o que lhe propomos aqui.

Além disso, o engenheiro de *software* poderá desenvolver o aplicativo de uma forma ainda mais robusta, permitindo que a imobiliária o alimente, periodicamente, com novos dados de apartamentos vendidos. Isso permitirá que o aplicativo se mantenha atualizado frente à evolução das condições de preços do mercado imobiliário, que podem subir ou descer com as flutuações da economia.

praticar

Vamos Praticar

Exemplo didático para regressão linear: como exemplo didático para a regressão linear, considere o proprietário de um restaurante que deseja aumentar as vendas investindo em propaganda na rádio da cidade. Considere também que o gasto nesse tipo de publicidade é calculado pelo número de inserções do anúncio na programação da rádio durante o mês. Com o cuidado de mensurar o efeito desses anúncios, o proprietário do restaurante somou, ao final dos meses em que fez o investimento com o anúncio, o número de vendas do prato filé à *parmegiana*.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. Rio de Janeiro: Elsevier, 2016.

O texto em referência descreve um problema de regressão linear, para o qual se obteve o seguinte modelo:

$$y = 117,38 + 9,62x$$

onde:

$$x = \text{número de inserções de anúncios durante o mês}$$

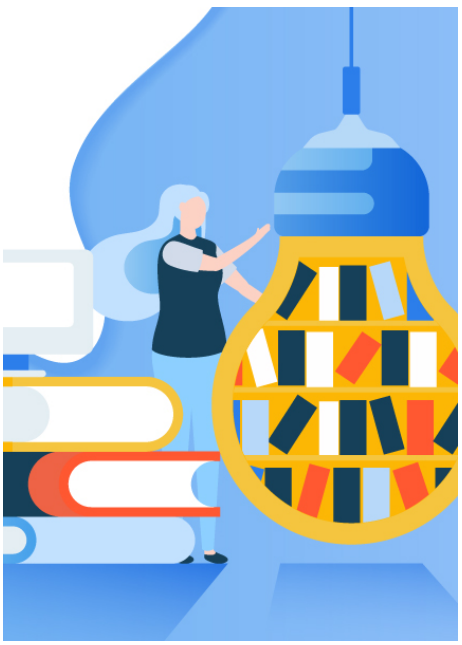
y = número de pratos de Filé a Parmegiana vendidos no mês

Para essa situação, assinale a alternativa correta:

- **a)** Os modelos de regressão linear são divididos em modelos de regressão linear simples e múltipla. O modelo desenvolvido para a situação aqui descrita é um modelo de regressão linear múltipla, onde há mais do que uma variável de entrada.
- **b)** A variável resposta deste modelo de regressão linear é o número de inserções de anúncios em um mês. A variável resposta também é chamada de variável independente ou regressora.
- **c)** O parâmetro 9,62 representa o número de pratos de filé à *parmegiana* que são vendidos em um mês em que não se fez nenhuma inserção de anúncios. Em outras palavras, representa o número de pratos y quando x = 0.
- **d)** Para saber quanto pratos de filé à *parmegiana* conseguirá vender no mês se investir em 50 inserções de anúncios, o proprietário substituiu o x da equação do modelo por 50 e obteve 598 pratos (arredondando para um número inteiro de pratos).
- **e)** Como a estatística e a ciência dos dados possuem em suas bases teorias matemáticas, não é possível aplicá-las a ciências sociais ou humanas (sociologia, história, antropologia, ciências políticas, direito, administração, filosofia, geografia, economia etc.).

indicações

Material Complementar



LIVRO

O Andar do Bêbado: Como o Acaso Determina Nossas Vidas

Editora: Jorge Zahar

Autor: Leonard Mlodinow

ISBN: 9788537801550

Comentário: este livro discorre sobre aleatoriedade, probabilidade e estatística. É um best-seller, que ficou vários anos como um dos mais vendidos na sua categoria. Dividido em 10 capítulos, usa de uma linguagem simples para nos contar, por meio de um passeio por vários casos interessantes, como o acaso determina nossas vidas.

WEB

Hans Rosling Mostra As Melhores Estatísticas Que Você Já Viu

Ano: 2006

Comentário: este TED TALK é, talvez, um dos mais representativos do famoso médico sueco Hans Rosling (1948 - 2017). Além de médico, Hans Rosling também era estatístico e orador. Dedicou parte de sua vida à difusão, por meio de inúmeras palestras e vídeos, da importância da aplicação da estatística ao estudo da saúde pública dos países no mundo.

ACESSAR

conclusão

Conclusão

Nessa unidade, contamos com a ajuda de dois personagens, uma corretora de imóveis e um estatístico, e pudemos ver – com essa valiosa ajuda – como é possível desenvolvermos uma capacidade preditiva se tivermos dados onde nos basear e modelos que "aprendem com os dados". Especificamente, iniciamos a nossa jornada por este mundo, o da "Estatística Aplicada à Ciência dos Dados", com os modelos preditivos chamados de regressão linear, simples e múltipla. Nas próximas unidades aprofundaremos nossa jornada por esse incrível e poderoso mundo. Vamos lá?

referências

Referências Bibliográficas

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. Rio de Janeiro: LTC, 2013.

OLIVEIRA, P. F. de; GUERRA, S.; MCDONNEL, R. **Ciência dos Dados com R**: Introdução. Brasília: IBPAD, 2018. Disponível em: https://cdr.ibpad.com.br/cdr-intro.pdf. Acesso em: 25 nov. 2019.

RITTER, M. do N.; THEY, N. H. **Introdução ao software estatístico R**. Imbé: CECLIMAR/UFRGS, 2019. Disponível em:

https://lume.ufrgs.br/bitstream/handle/10183/188778/001087242.pdf?sequence=1&isAllowed=y. Acesso em: 4 dez. 2019.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. Rio de Janeiro: Elsevier, 2016.

WICKHAM, H.; GROLEMUND, G. **R for data science**: import, tidy, transform, visualize, and model data. Sebastopol: O'Reilly Media, 2017.